Information Services
[redacted]

# IS

**RESEARCH NOTE**

*Strategic Planning & Research Advisory*

By: T.M. Ciesla
*ISRN - 3*

January 31, 1993 Rev. 1

## ELECTRONIC DOCUMENT MANAGEMENT SYSTEMS (EDMS)

**Installation of an EDMS at [redacted] Park would satisfy the information distribution requirements of OSHA's 1910 regulation and serve as a pilot program for a Houston wide EDMS. This *Research Note* explores the technologies required to implement this type of system, and the challenges that it will create.**

## INTRODUCTION

***Key Issue***

Examine the technologies and issues related to introducing an EDMS to the Solvay culture.

***Assumptions***

A stable, robust LAN backbone exists to service all users in the pilot project.

The document management products selected for the pilot project will be capable of eventually supporting an enterprise wide EDMS.

The release of an RFP will be an integral component of any pilot. This underscores the knowledge that a degree of outside expertise will be required to integrate the various technologies that make up the EDMS.

The data residing in corporate databases typically represents less than 20 percent of a company's total information resource. A large portion of the remaining data is captured in a variety of text and image based documents, ranging from memo's to internal research findings. While traditional records management attempts to track this data via centralized filing systems, often it is difficult to retrieve information that is pertinent *and* represents a complete collection. This problem is inherent with information distillation using the *'point-of-entry'* style of inputting data into a system, using keywords or subject headings, etc. The true value of a document is unknown at any given point in time, the value is constantly changing due to external world events. For example, four years ago, no one would have used 'political' as a keyword for an article on Ross Perot. External events changed the value of that article over time.

Electronic document management systems allow access to all documents, in their original form, on-line, all the time. Queries can search documents for any variety of text or text combinations, depending on the research need at the time. This *'point-of-research'* type of distillation, maintains the original document content while allowing for infinite ways of accessing the data within. By incorporating hypertext into the documents, links can be made to other documents, images or drawings, which can be accessed by the click of a mouse button.

With an EDMS, even the definition of the word 'document' changes. A document is no longer one or more pages, it is a collection of data that can be dynamically categorized for the purpose of organization, ease of understanding, and accuracy. Documents may now contain text as well as images, sound or video, when multimedia is introduced. Furthering this paradigm shift is the concept of the 'virtual document', where documents are created on the fly by pulling pertinent data from

other resources within the system.

In short, electronic document management changes how we store and retrieve documents, as well as how we interpret and distribute the information contained in those documents.

## WHAT IS EDMS?

Simply defined, electronic document management is the systematic and automated organization of documents through some type of database or equivalent. While often referred to as a 'technology', an EDMS is actually an integrated collection of divergent technologies which allow for the collection, storage, retrieval and manipulation of data stored in a variety of formats. ***Imaging*** is used to convert paper based documents, photographs or drawings to an electronic format which can be stored in a database. ***OCR scanning*** is intelligent imaging that indexes all or some of the text of a legacy document being converted to electronic format. Once converted in this manner, these documents may be searched for specific text.

***EDMS Challenge***

Knowledge, so central to innovation must be distributed throughout an organization, not concentrated in any one area of a company. Sharing this knowledge, which exists as text, graphics, sound and video, is the challenge of an electronic document management system.

***Text retrieval*** allows the storage and retrieval of coded documents, that contain text, based on their explicit or implicit content such as words, sentences, phrases or concepts. Most major text retrieval packages now include Hypertext capabilities. ***Hypertext*** is a methodology that focuses as much on organization as it does retrieval. By electronically linking nodes of related textual information, it is the only text search methodology that mimics the mind's associative memory process. Windowed links are used to access paths through a text database, providing a user with the ability to explore all available information on a given topic, across document boundaries. The hypertext concept can be extended to include the linking of text to other forms of information (i.e. image, audio, graphics, and video). This technology adjunct is known as ***hypermedia.***

Together, imaging and text retrieval may be viewed as the core services of an electronic document management system. Each represents a mature technology with a track record in the business community. While these core services may be traced back almost twenty years , some newer 'enabling' technologies are being brought under the document management umbrella.

***Multimedia*** is high bandwidth communication which allows parallel presentation of information, potentially using all senses simultaneously. Its rise in popularity can be attributed to the spread of ***Graphical User Interfaces (GUI's),*** of which, the Windows package from Microsoft is on the verge of becoming the *defacto standard.* Through the use of third party vendor applications supporting Windows ***dynamic data exchange (DDE),*** powerful front end tools can be quickly developed to work within the EDMS arena. Most imaging and text retrieval vendors now include DDE support, which enables the user to

transparently copy data from one application to another. This capability is important if users want to integrate imaging features with other applications.

The key to implementing a successful EDMS is to integrate these various technologies into one seamless application. We will now examine each of these technologies in detail to discuss the variety of options and standards that control each technology and impact their integration.

## ELEMENTS OF AN EDMS

### *IMAGING*

The role of imaging systems in an EDMS will vary based on the focus of a given management system. For example, in [redacted] Pilot Project, imaging will primarily used for scanning in a limited number of documents (MSDS's, photographs, etc.). In an enterprise wide application serving the Solvay Houston group, imaging would play a central role in moving to a paperless office.

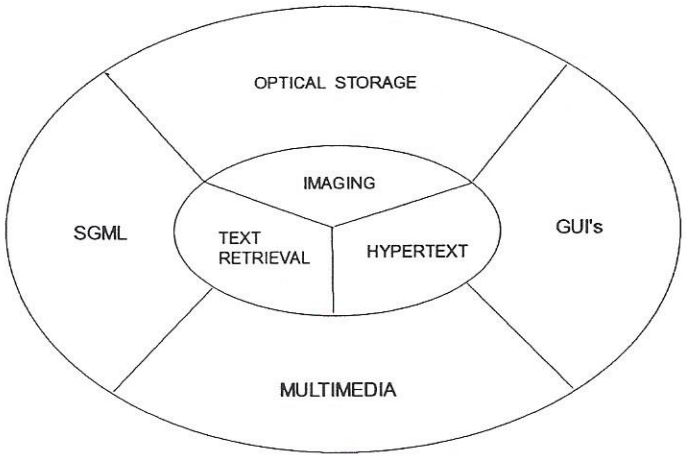



***Figure 1 Key Elements Of An EDMS***

SCANNERS

Regardless of size, imaging systems contain the same basic components: scanners, enhanced display monitors, image software, storage devices, servers, and printers.

*What is a RAID?*

RAID stands for '*Redundant Array of Inexpensive Disks*'. RAID technology provides an array of drives under the control of one device driver. Data is written to the drives as though they were one system. Redundant writes prevent data loss if one drive fails.

Total storage capacity achievable with RAID systems exceeds 60 GB.

A minimum scanner configuration would be a 300 DPI flatbed model with interface card and auto document feeder. File compression (software or hardware based) should support CCITT group 3/Group 4, which equates to a 10 - to - 1 compression rate. OCR Scanners should have a 90 - 95% accuracy rate.

MONITORS

To use imaging effectively, that is to encourage use of the viewing station versus printing documents, the viewing station should be a 19-21" enhanced resolution monitor as a minimum. Stepping down from a high resolution monitor to a SVGA monitor results in a 30% decrease in image resolution. Stepping down from an SVGA to a monochrome VGA results in an additional 30% decrease in resolution. The 17" monitor, an intermediate solution, will average $1700 (plus video card) while a 21" monitor will average $3500 (plus video card).

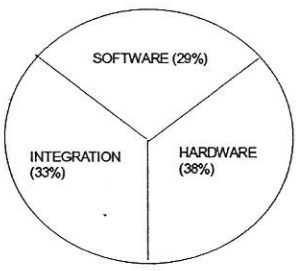
SOFTWARE

Imaging software should support multiple platforms; image scanner, OCR, and CAD input; as well as CCITT Group 3/Group 4, TIFF and PCX file formats. To ensure future data portability the system should support a major relational database package (Oracle, Sybase, Informix), or be able to export to one at a minimum.

STORAGE

A driving force behind the expansion of imaging is low cost and efficient storage. Scanned images can be stored on high volume optical disks or magnetic disk. Optical disks are available in write-once-read-many (WORM) and Erasable formats. WORMS are best suited for archival purposes or where there is a legal requirement to protect data integrity (i.e. MSDS's). Each type of drive is capable of being used in an optical Jukebox system, which is capable of storing multiple disks through robotic selection.

While optical disks are cost effective ( .05¢-.50¢/MB media only vs $2-$5/MB for Hard Drives), access times are slower than magnetic hard drives (40-200 ms for WORMS vs 1-15 ms for magnetic). The cost of a jukebox style optical system averages about $1.50/MB. A 6-10 GB Jukebox will cost from $10,000 - $15,000 , while a 4GB RAID hard drive system (with redundant disk) will average about $19,000.




*Costs Of Imaging*

## *TEXT RETRIEVAL*

At a minimum, a text retrieval system must provide the ability to input text documents, establish search methodologies, query the database and output results. The basic architecture of a text retrieval system is analogous to that of a traditional DBMS. The major difference is the inclusion of a text field containing the name of the textual document and a pointer to the document file.

The textual documents can be stored in the database, or at the operating system level in a native file format. Storing the documents in the database provides greater control and typically decreases storage overhead. Storing the documents in their native format offers greater flexibility for editing and sharing the documents with other applications.

The search and retrieval of documents in a database can be done through several methodologies, which are detailed in figure 3 below. Some vendors rely exclusively on one particular method, though it is common to integrate several methodologies utilizing the benefits of each where appropriate.

***Precision & Recall***

The effectiveness of a text retrieval system lies in the ability to control both Precision and Recall. ***Precision*** asks the question: "Is what I found what I was asking for?" ***Recall*** asks the question: "How much is out there, and how do I know I have found it all?"

| SEARCH METHODOLOGIES | | |
|---|---|---|
| **METHODOLOGY** | **BENEFITS** | **DRAWBACKS** |
| ***Inverted Positional Index***<br>Indexes every word in a document alphabetically and stores its location within the document | - Minimal query I/O<br>- Retrieval rate minimallly impacted by database size. | - Index overhead (30 - 150% size of text)<br>- index maintenance<br>- Stopword maintenance |
| ***Inverted Non-Positional Index***<br>Similar to above, but tracks the occurence of words at the document level. Details of word positions are not tracked. | - Minimizes index overhead (30% of text)<br>- Minimal index maintenance | - May require alternative search method for complex query.<br>- Requires document access for complex quiery. |
| ***Free Text Scan***<br>This methodology employees no indexing. The system reads through every document for the query term. Usually requires extra hardware to expedite the search. | - No index overhead or maintenance<br>- No stopwords | - Document i/o<br>- Potential hardware solution |
| ***N - Gram Array***<br>Similar to inverted index approach. An alphabetical index is created of all unique suffixes; unigrams, duograms, trigrams | - Provides efficient suffix wildcarding<br>- Enables sophisticated lexical analysis. | - Indexing overhead (50-80% of text)<br>- Index maintenance. |
| ***Pattern Recognition***<br>Words in the text and the query are reduced to unique ASCII character patterns. Query patterns are compared to the words patterns for simularity. | - Basedon neural technology<br>- Allows "fuzzy searching"; mispelling tolerance.<br>- User controllable retrieval speed, spelling accuracy. | - Index maintenance and overhead.<br>- "Fuzzy" search can adversely effect precision.<br>- Exact matches require alternate technology. |
| ***Concept Based Clustering***<br>Subject profiles for each document is used to position document in a storage "area". Concept-based clustering provides the ability for capturing human judgement and explicit knowledge. | - Effective for very large and/or specialized databases.<br>- Innately provides conceptual searches<br>- Can identify similarities between documents. | - High implementation costs.<br>- Clustering maintenance.<br>- Requires large document collection |
| ***Statistic Based CLustering***<br>Based on relationships of words to concepts, the system creates a virtual storage area of 'n' dimensions. Documents are virtually stored or clustered in this area. | - Effective for very large and/or specialized databases<br>- natural language query<br>- Innately provides conceptual searches. | - Reliance on statistics<br>- Clustering maintenance<br>- Requires large document collection |
| ***Hypertext***<br>Electronically links nodes of information within a document or between other documents through a text database. | - Facilitates capture of document inter-dependence<br>- Predefines links and paths allow for easy version control.<br>- Can be intergrated with other methodologies. | - Requires special skill set for authoring.<br>- Potential for creating maze of links.<br>- Lack of stanadards controlling navigation. |

***Figure 3. Comparing Search Methodologies***

## *HYPERTEXT*

As one of the more revolutionary methodologies, HyperText deserves special mention. It is the only text search based on the associative memory process, allowing a user to walk through a series of informational doorways and gather knowledge through association. Hypertext can be used to control access through a single document, or through several forms of information. This facilitates the creation and maintenance of multiple versions of documents.

While Hypertext assists in the capture of document inter-dependence and predefined links allow for version control, one must use caution when implementing this system. Developing hypertext links requires a special type of skill set and an objective approach to effectively author without creating a maze of links.

With ***Hyperlinks,*** a user may move from one document to another by simply clicking on a highlighted word or graphic. For example: clicking on a symbol for a reactor on a P & ID might call up the maintenance procedure for that unit.

A ***Hyperlinked*** site plan could be used as a data navigation tool for a specific control room's documents by simply clicking on that control room building.

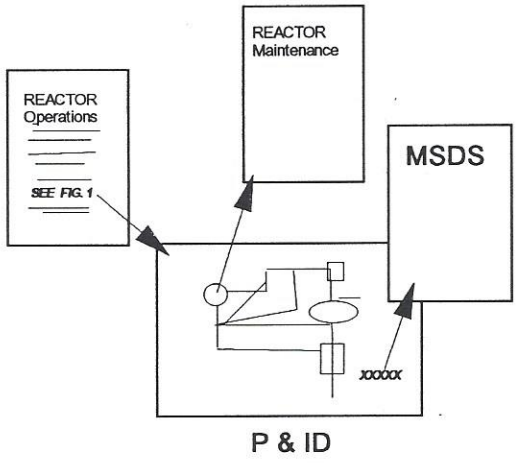



***Figure 4.*** *Hyperlinks Between Documents*

## *SGML*

"*Standard Generalized Markup Language*", ISO Standard 8879, is a standard for the structural markup of text documents. SGML has been adopted by the Department Of Defense for the multi-billion dollar CALS Initiative, the European Economic Community, and many private organizations.

Structural markup incorporates the use of tags which are used to define sections of a document (i.e. Title, Author, Date, Paragraph, etc.). SGML is a complex markup system which specifies a formal, machine processable system for tagging and verifying text structures. By indexing on the structure as well as the text, searches can be performed on particular structures in a text, or on combinations of structures.

While SGML affords effective support for Hypertext, Multimedia, and enables easy integration of text databases, the cost of implementation has slowed adoption. Editing a large inventory of existing text files to include a few structural elements is useful but labor intensive. Many companies are positioning themselves for the future by requiring SGML capability from the application purchased, without having immediate plans to implement it.

***On-Screen Appeal***

Structure can be lost as documents are converted from authored format to EDMS format. If preserving the ***visual metaphor*** (i.e. viewing what was a hardcopy '*page*' as *a 'page*' on a monitor), is important to the users, research the EDMS product in detail. Not all vendors provide this function.

### *MULTIMEDIA*

Computer based Multimedia integrates text and numbers with sensory information in its original undisturbed form, engaging the senses of sight and sound. The biggest advantage of multimedia is the level of interaction it affords the user, which influences how much information is retained. According to a Stanford University study, people remember 20% of what they hear, 30% of what they see, and 60% of what they interact with.

In the EDMS arena however, Multimedia is just now taking hold, with major implementation approximately two years away. Like SGML, Multimedia should be supported by the application selected, but the impact to the organization is a few years into the future.

## IMPLEMENTING AN EDMS

Accepting any single new technology into a company's established infrastructure is difficult for most companies. Adopting the variety of technologies associated with an EDMS increases the cultural resistance because of the complexity. The key to a successful EDMS implementation is doing the up front planning and analysis needed to define the architecture of the infrastructure, the EDMS goals, and the document processing and distribution. A company can expect to spend about a tenth of the total system cost on this analysis.

### *THE ARCHITECTURE*

The impact of an EDMS on a company's existing infrastructure will be dependent on the scope of the implementation and the approach used for selecting an EDMS product (or products). Some products are platform-specific and/or database specific. The majority of vendors,

however, support Novell, Banyan, TCP/IP, DOS, Windows, and Client/Server for distribution platforms; and Oracle, Sybase, and Informix as common databases.

The impact of an EDMS on a LAN backbone will depend on: (1 the amount of imaging used, (2) the size of the image documents, (3) the type of compression used, (4) the distribution of databases and indexes, (5) the frequency that the EDMS will be accessed, and (6) the other applications and overhead traffic on the LAN.

Typically an EDMS will require a application server, as well as a image database server and a text database server ( potentially with a separate indices server). Storage is typically accomplished via WORM drives (stand alone or Jukebox), RAIDS, or a combination of these (See Figure 5).

*Use Of Open Systems*

Because of the rapid changes occurring in the EDMS technologies and the desktop arena, expect whatever you select today to have a two year lifetime. The only way to protect against rapid obsolescence is to incorporate products that are as open as possible, or can at least be exported or upgraded easily. *Avoid proprietary technology whenever possible.*

*ViewStation Profile*

The typical Viewstation is a 486 desktop PC with a 19 -21", high resolution monitor, 8-10 MB of RAM, and internal hard drive. A soundboard is optional.

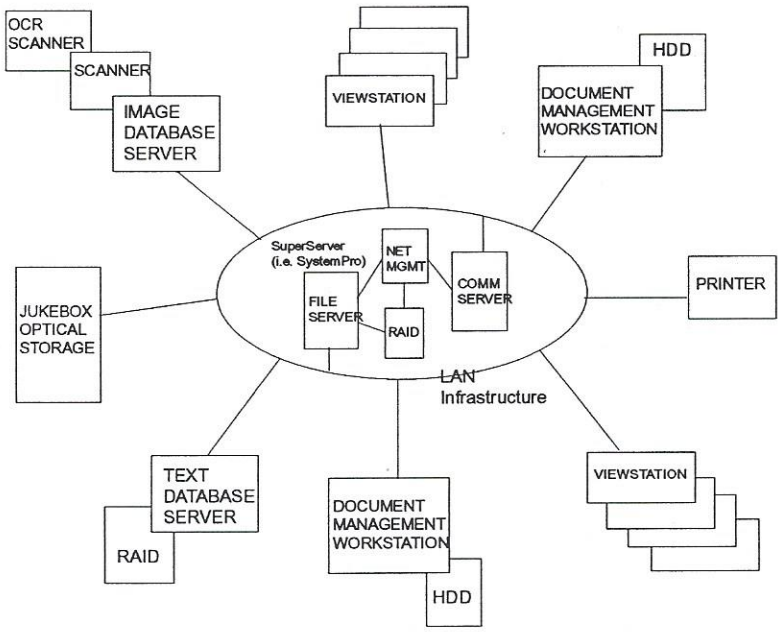



***Figure 5. A Generic EDMS Scheme***

## *THE DOCUMENT SURVEY*

Each document that will reside in the EDMS system must be reviewed against three key phases of the document life cycle: (1) Document creation, (2) Document activity, and (3) Document archival. The lack of this fundamental understanding - as well as the size of the document

population - makes the development of a successful solution strategy impossible.

One good evaluation tool is known as the '***EDMS document worksheet***' which is used to record specific information on each document category or type that will be used. At a minimum, this worksheet should capture:

- ***format*** (memo, letter, graphic, etc.)
- ***media*** (paper, electronic)
- ***retrieval frequency***
- ***concurrent usage***
- ***archival format & legality***
- ***archival access***

This information can then be used as a common framework for users, evaluators, and implementors; and identify major areas of benefits and concerns. It will also help determine the degree of imaging needed (and therefore the number of scanners and OCR devices), the amount of document re-authoring to be done, and the amount of on-line storage capacity required. Finally, the survey will serve as part of the foundation for the final RFP and/or project plan.

## *SYSTEM CONFIGURATION/ADMINISTRATION*

In the context of this report, the term system configuration encompasses: the setup of thesaurus definitions, document re-authoring procedures, hypertext standards, document naming standards, configuration control procedures, and database/index distribution.

In addition to EDMS-specific hardware, software and interfaces (i.e. device drivers, GUI's, TCP/IP, etc), the project implementation plan should include sufficient resources and time for converting paper and electronic documents into EDMS format. Costs of imaging legacy documents range from 50¢ to $5 a page, depending on the quality of the documents. Costs of re-authoring electronic documents are difficult to quantify- being heavily dependent on the level of linking and structural changes desired.

For example, in large documents (15 pages or more?), how will the user navigate through it? By page number? Through a hypertext linked table of contents? By hypertext linked chapter headings? Hypertext within the body of the document will impose additional man hour requirements of a knowledge worker familiar with the subject matter.

An approach used by many companies is to minimize the use of hypertext links during the initial implementation phase if resources are limited. These links can always be added after the pilot is successfully operational.

At the system level, how will the user navigate through the system? Products which lack an easy-to-use GUI interface may overwhelm a user and appear to be something designed for NASA. If menu's serve as the navigation tool, expect to spend a substantial amount of time developing hierarchial trees. The menus themselves may be centered around functional processes, business centers, corporate subsidiaries, or document groupings. Ultimately, the final choice of a navigation method should be made with heavy user participation.

Ongoing administration of an EDMS will involve: maintenance of databases and indexes, monitoring storage capacity and backups, management of incoming documents, and enforcement of security issues. For the majority of EDMS's, some level of permanent staffing will be required to monitor the various technologies and issues involved.

## *EDUCATION*

An EDMS represents a dramatically different way of working with information; so it follows that it must be understood to be used effectively. A good indicator of failure is when such a system generates ***more*** paper. As such, training - essential to the projects' success - will take many forms, touching all levels of a company.

**Implementation Team** ------ with minimal IS representation, this team is a cross section of the enterprise. Early education of the team members is essential for two reasons: (1) training will develop familiarity which breeds comfort and confidence, and (2) informed members will be equipped to handle the curiosity of management and users, serving as champions of the project.

**Workgroups** ----- while management may sanction the introduction of an EDMS, it is the users who ultimately determine its success. Three sub groups make up this level: the 'view-only' user, the 'knowledge-editor', and the EDMS administrator.

> The ***view-only user*** would typically be a control room operator who only wants to view a procedure, a drawing, or a safety data sheet, and possibly attach an electronic 'post-it' note. Training at this level would include a general understanding of the interfaces and navigation tools.
>
> The ***knowledge-editor*** is a professional such as an engineer having Edit rights granted to alter procedures and/or drawings in the system. This level of training includes that described above, plus an understanding of the editing tools and management of change issues.

The ***EDMS administrator*** must have a thorough knowledge of all the applications on the system and how they are integrated. This training is normally done by the product vendor or system integrator who does the installation.

**Information Services** ---- IS staff members must become experts in the technologies, implementation strategies, and support issues of the EDMS. Training at this level will be on-going - plan for attendance at industry seminars and conferences in your budget from the outset. These are cost effective ways to keep abreast of the most current issues and technology changes.

**Management** ----- having an organizational sponsor with adequate clout is half the battle of successfully installing the EDMS. Management should have a basic understanding of how the system works and the business value it offers. This education will help make long term commitments to the system and ride out the short term obstacles that may arise.

## *WORKFLOW*

Workflow automates, coordinates and tracks business processes and procedures by proactively integrating and managing existing documents and applications. It is a minimalistic approach to technology - improving what already exists rather than layering more hardware onto existing processes.

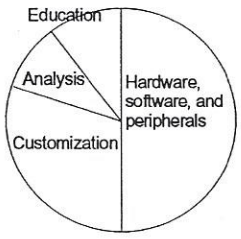



***Figure 6.***
***The Real Cost of EDMS***

In the mid - 1980's, workflow software application were developed by companies such as FileNet with their *Workflo* product, serving primarily image based applications. More recently Lotus Notes has gained increasing popularity in some vertical markets, based on it's workflow management capabilities.

In terms of EDMS implementation, performing workflow analysis for the departments involved is critical for the success of the document management project. Knowledge of document routing, task assignment, electronic signatures requirements and configuration control is necessary for the proper design of an EDMS. Whether the workflow study is done in-house or through an EDMS integrator will depend on available personnel resources in a given organization and the size of the EDMS project budget.

## JUSTIFYING THE COST

According to a study conducted by the Delphi Consulting group of Boston, over the next two years, a Fortune 1000 company can expect to spend approximately $500,000 on an EDMS. Of that total only 50 percent will be spent on hardware, software, and peripherals. An equal amount will go towards analysis, implementation and education. Despite these costs, 90 percent of companies who have had an EDMS during the past 3 years believe it is a necessary component of their integrated applications environment. Many maintain it has given them the competitive edge promised by numerous other new technologies.

To determine the value added potential of any technology, many organizations feel that the only measure that means anything is the Return On Investment (ROI) calculation. Other methods try to blend those hard cost considerations with assessments of how well different technologies support the groups business goals, such as time-to-market, or customer service. Proponents of these methods say they more accurately reflect the organizations priorities.

*Need More Information?*

If you would like to obtain further information on issues discussed in this report, please contact the author at [redacted] facility.

However, the strategic factors that are important to an enterprise are rarely ROI. It is often impossible to put a cash value on the effect of, say, improving the information on which management bases its decisions, faster retrieval of documents, or more effective research results.

While increased productivity, communication, and elimination of paper are often listed as key benefits of an EDMS, the true value to a given organization will vary depending on the reason for the implementation. For example, the [redacted] pilot project is being considered, in part, to satisfy the information distribution requirements of OSHA regulation 1910, and in part, to assist in the long term strategic plan of building a world class IS system.

Granted, that while a successful pilot at [redacted] ( a project with a very specific purpose), would hopefully obviate potential penalties, that reason alone is a questionable way of justifiying cost. The true value of this pilot is ultimately a subjective assessment, and one that will be realized over a period of time. 'Just-in-time' information supplied to a control room operator may save lives or prevent the waste of raw materials, but the *true value* of that service at any given moment is intangible.

## CLOSING REMARKS

In a time of global competition, information is now regarded as a strategic weapon and the average employee has been transformed into a knowledge worker. When you consider that knowledge workers spend between 40 to 60 percent of their time working with documents, it's clear that document management will play a crucial role in

increased productivity. For most organizations, the cost of handling documents is second only to payroll.

Separating the myth from the reality is sometimes the hardest part of evaluating new products and systems. As a nascent *'technology'*, Electronic Document Management exemplifies that challenge to many IS managers. Remember that an EDMS cannot be implemented overnight. Planners should assume a two year horizon for completion. All avenues of vendors, the media, consultants, and users with existing systems, should be reviewed. Keep in mind that this is a rapidly changing technology that demands ongoing education.

While there is no substitute for a solid education in all the technologies that make up an EDMS, knowing the right questions to ask is half the battle.