



Document Search Methodologies

Different Approaches to Document Retrieval in Electronic Document
Management Systems

Jon Clark, Vice President Sales, Cabinet NG, Inc.

6/1/2009

Document Search Methodologies

The purpose of this white paper is to take a look at various ways to search for and retrieve business documents. Today's Electronic Document Management Systems (EDMS) have proven that search and retrieval efficiency can be improved dramatically over manual, paper-based filing systems.

Most EDMS use one or more of the following:

- structural search
- keyword and/or metadata search
- full text search

Each of these methods has advantages and disadvantages depending on the specific business application. This paper explains each of these methods and presents the advantages and disadvantages of each.

We will also explore some of the key considerations for you to be aware of as you think about how EDMS search and retrieval would work best for your company.

Let's Play a Game of 'Find the Folder'

Before we get in too deep with the technology side of searching for documents, it is important to consider how businesses perform routine document searches. Most businesses today store documents across three or four separate locations: centralized paper-based file cabinets, paper-based file cabinets in employee's offices, folders on shared server drives, and local desktop hard drives. Most companies try to centralize all filing to a single file room and a server, eliminating storing of business documents on local hard drives and file cabinets in people's offices. This is usually the first step toward making the transition to an EDMS.

For example, the HR manager (Pat) receives an urgent inquiry from an insurance company requesting a document from a specific employee's records. Here is the process:

1. Pat goes to the personnel file cabinet.
2. Pat has the key to the cabinet, so Pat unlocks the cabinet and looks for the employee folder.
3. Assuming the last person to use that folder replaced it in the correct location, Pat pulls that folder.
4. Pat locks the file cabinet.
5. Pat takes the folder to the copy room and faxes the necessary document to the insurance company.
6. Pat returns to the personnel file cabinet.
7. Pat unlocks the file cabinet and re-files the folder in the correct location.
8. Pat locks the file cabinet.

9. Pat returns to office and resumes work.

If this sounds like a dream sequence from a Hollywood film, then you have already figured out that in real life, the retrieval process never goes this smoothly. There are countless speed bumps that inevitably appear throughout this process. And it starts before we even get to Step 1. Pat gets another call or someone walks into Pat's office and the trip to the file cabinet gets delayed, if not forgotten.

What are some other speed bumps? Here are a few of the classics:

1. Pat can't find the key.
2. Pat spends time searching for the folder and can't find it. Someone else has it out or has misfiled it.
3. Pat runs into Sandy and they have a cup of coffee and discuss important business matters and maybe the movie they saw over the weekend.
4. Pat lays the folder down in the copy room and gets distracted. A disgruntled employee takes the folder and discovers private information. Mayhem and legal action ensues.
5. Pat makes it back to the file cabinet unscathed, but accidentally misfiles the employee folder, making it impossible for the next person to find.
6. Pat forgets the original objective of the trip to the file cabinet and decides to go have a smoke.

Is this an exaggeration? Probably not. You have most likely seen this happen. As an isolated event, it doesn't seem like that big of a deal (except the legal part). But it is not an isolated event. This is a process that is repeated by multiple people across your organization every day. Even in the perfect scenario where everything is always filed correctly and in a timely manner, too much valuable time is wasted filing and retrieving documents.

This is why there are many EDMS products and value added resellers who are dedicated to helping businesses run more efficiently. EDMS fixes the inefficiency inherent to paper-based filing systems. And the biggest financial impact on businesses that adopt EDMS is reduced labor cost. Time spent filing, retrieving and re-creating and processing documents is typically reduced by 50% or more. The biggest time savings of all comes from retrieval (search) time.

Structural Search

Structural search is the method most closely related to the file cabinet, file folder approach because it relies on a consistent, hierarchical and controlled structure for storing documents. The best systems are able to emulate the physical filing world to make the transition to EDMS more seamless for people who are accustomed to working with file cabinets, folders, index tabs, documents and even paper-clipped or stapled documents. So now Pat can log in to the Personnel cabinet, find and open a folder, identify the required document and even email, fax or print the document while the insurance company representative is still on the phone.

Structural search is highly dependent on the Graphical User Interface (GUI) used by the EDMS. A good GUI is intuitive and requires little or no training to search for and retrieve documents. If a user can

visualize the filing structure and navigate to specific documents with minimum clicks and data entry, the GUI is probably most responsible.

It is important that a structural search provides similar control over access to documents. In the personnel example, only people with the key to the file cabinet have access to employee files. (Unless a file is accidentally left lying around.) A good EDMS should expand on the 'key to the cabinet' security model and allow access control at the folder and document level. This allows people instant access to only those documents they have permission to use.

The structural search is the most efficient of all the search methodologies because users can quickly get to the one specific document that is sought with a minimum of clicks and keystrokes. Other methods provide a *narrowing down* type approach that requires sorting through a list of 'matches' to find the exact document you are looking for. The structural search is most efficient when used within a system that allows users to preview documents without actually opening the document. For example, if there are three invoices that have been scanned and filed and indexed similarly, it saves a lot of time if you can preview the three documents without opening them to help determine which document is the one you are after.

The main disadvantage of the structural search is that it can take a little more time to file a document into the correct location. Every document needs to be filed into a specific cabinet and folder. Most systems that offer structural search have tools for filing and creating documents as efficiently as possible to minimize filing time. And keep in mind, scanning a document and picking a cabinet and folder on your desktop display is still far faster than filing paper documents into file cabinets.

Not all systems provide true structural search capabilities. Look for an interface that is folder-centric in systems that do structural search well.

Keyword Search

Keyword search can be used in conjunction with structural search or as a standalone search method. Dictionary.com defines 'keyword' as:

key·word

1. a word that serves as a key, as to the meaning of another word, a sentence, passage, or the like: *Search the database for the keyword "Ireland."*
2. a word used to encipher or decipher a cryptogram, as a pattern for a transposition procedure or the basis for a complex substitution.
3. Also called **catchword**. *Library Science*. a significant or memorable word or term in the title, abstract, or text of an item being indexed, used as the index entry.

From <http://www.dictionary.com>

In our case, the third definition applies. An EDMS allows users to index documents with keywords. Those keywords can be entered later in a search field and a list of documents associated with a keyword or set of keywords will be presented. The more keywords associated with a document the more specific the searches that can be performed. The fewer keywords, the more likely you would receive a longer list of documents returned by the search.

Some EMDS require manual entry of all keywords, and other systems automate keyword input. The best way to use keywords is with systems that allow manual and automatic keyword entry. For example, let's say Pat, our busy HR manager, needs to scan and file a Vacation Approval form for an employee named Kelly.

If Pat can go to Kelly's folder in the EMDS and scan the document, then information like Kelly's name and employee number as well as keywords like 'vacation' and 'request' can be automatically attached to the document in many good EDMS. In addition, Pat may want to add keywords to the document to further assist with possible future searches. Examples might be keywords like: 'approved,' 'personal,' 'comp time,' etc. Now Pat has a couple of ways to find this document in the future. A structural search would allow Pat to look in Kelly's folder for documents with the title 'Vacation Approval.' Or Pat could do a keyword search for documents with the keywords 'Kelly,' and 'Vacation.'

Or better yet, if the EDMS allows for a combination structural and keyword search, Pat could navigate to Kelly's folder and do a keyword search for 'Vacation.' This type of flexibility gives users options on finding the most efficient path to a specific document.

The biggest advantage of using keywords to search for documents is that it gives the user the most freedom to index and perform searches. The biggest disadvantage of using keywords to search for documents is that it gives the user the most freedom to index and perform searches. An approach to EDMS that give each user too much freedom to index documents can be extremely risky because no two people will ever use the same indexing convention. Keyword indexing and search is most efficient when used within the context of a structural search. EDMS that rely solely on keyword indexing and search are less efficient because every single search performed requires input of keyword search criteria. Additionally, most keyword searches will return multiple document results, requiring additional filtering by adding even more keywords to the search criteria. Keyword searching also does not guarantee you will find the document you are looking for. If the person that originally filed the document added no keywords, then a keyword search will not find the document. This is why it is vital that the EDMS should have multiple search capabilities.

Metadata

In addition to keywords, many EDMS allow the use of metadata for performing document searches. Wikipedia defines metadata as:

“Metadata (meta data, or sometimes metainformation) is "data about other data.” An item of metadata may describe an individual datum, or content item, or a collection of data including multiple content items and hierarchical levels, for example a database schema. In data processing, metadata provides information about, or documentation of, other data managed within an application or environment. This commonly defines the structure or schema of the primary data. The term should be used with caution as all data is about something, and is therefore ‘metadata’ in a sense, and vice versa.”

EDMS that support metadata searching have certain metadata associated with documents. Examples could be information like: document creation date, document edit data, who created the document, who viewed the document last, etc. The real power of metadata is when used in conjunction with keywords. By adding this capability, Pat could go to Kelly’s electronic folder and run a search on keyword ‘Vacation’ with a create date of ‘May 21, 2009’ to find the exact document required.

The advantage to using metadata for document searches is the automatic way metadata is typically applied to documents. The EDMS applies metadata like ‘create date,’ ‘created by,’ ‘edit date,’ etc. without the need for any user intervention. And since metadata is applied systematically, it is consistent and reliable across the EDMS, no matter how many people are indexing and searching. The disadvantage of metadata is that it is limited to only certain types of document information; therefore it is not very useful unless combined with other search methods, like structural or keyword.

Full Text Search

Full Text Search (FTS) is yet another way to search for a document. FTS involves looking for a document based on a word or phrase that may be contained within. For example, if Pat wants to find all documents in the electronic HR cabinet which contain the phrase ‘law suit’ FTS would be very useful.

FTS first and foremost requires that documents contain text. An EDMS that provides full text search indexes the text contained in all the documents within a database. Depending on the size of the document repository the FTS database can become fairly large. As documents are filed and indexed, they become *searchable* using the EDMS FTS feature. This is a straightforward process for documents like emails, MS Word®, MS Excel® and other text-based documents. However, scanned documents do not contain text (a scanned document is an image) so they must be converted to a format that contains text and is searchable.

Today’s EDMS typically provide an option to do this conversion for you. This is usually referred to as OCR conversion. OCR stands for Optical Character Recognition which is the technology that is used to convert an image to text. (The standard format that is created is a searchable PDF.)

The process of running an FTS is very simple with most EDMS. The user types in the word or phrase and the system does a search and returns a list of ‘hits.’ Most EDMS that support FTS will also let the user run a Boolean search which allow for the use of ‘and’ ‘or’ type criteria, wildcard searching using the * as a wild card character and even fuzzy searching which will find words that match closely if not exactly to the text entered by the user.

According to Wikipedia, today's OCR technology is greater than 99% accurate. Source: http://en.wikipedia.org/wiki/Optical_character_recognition. Let's say OCR is 99.9% accurate (which it is not proven to be). That sounds pretty good doesn't it? It is. The problem is that if you rely solely on OCR conversion and FTS for searching you would be guaranteed to lose 1 out of 1000 documents. Most businesses cannot afford that.

There are a couple of disadvantages to FTS. First, the extra step required to convert scanned images to a text searchable format using OCR takes a little more time and adds to the cost of the system, and the conversion of that text is not 100% accurate. The second disadvantage is that FTS, like keywords, should not be relied upon as the sole search method. If Pat runs an FTS on the HR cabinet for 'law suit' we would hope there would not be a long list of documents presented. But if Pat runs an FTS on 'vacation request' there could be 1000's of hits.

However, the power of FTS to locate documents based on their content has many applications. Therefore, when used in conjunction with structural and keyword search methods, FTS is a great additional feature to round out a complete EDMS.

Filing vs. Retrieval

Taking a few seconds to index and file a document correctly could save you and others many hours of retrieval time. The less time and effort you put into indexing and filing a document, the harder it will be to find. By the same token, a good EDMS should provide streamlined approaches to automating the indexing process as much as possible. Efficient and high quality filing and indexing into an EDMS will result in efficient retrieval.

There is no magic button that you can press on your scanner to automatically file all documents correctly so they will be easily retrieved. But this should not deter you from exploring EDMS to make your business more efficient. The productivity improvement will be realized by every person in your business that files, retrieves, creates, edits and processes documents.

Conclusion

As you may have already concluded, no one search method is ideal for every business and every application. As with most things, there are compromises to be considered. For example, going with the easiest way to get files into an EDMS might mean it will take a little longer to retrieve. Careful, detailed indexing of every document may make it easier to find later, but there is a cost associated with this approach due to more time spent up front filing documents. The best solution is to have the ability to run any of the discussed searches depending on the business process and the most efficient overall approach.