White Paper: Document Indexing Exposed

Document Indexing Exposed

by Joe Hill

All of us, at one time or another, have been frustrated looking for but not being able to find documents that we know we filed in a safe place in our filing cabinet or on our computer system. Or maybe you have experienced the frustration of looking for a piece of data or information that you wanted to recover or reference and couldn't because you could not recall where the document, article, or file was located. Through using effective document indexing in a content management system, those frustrations can be largely eliminated. But there is more behind the curtain than one may first imagine and the purpose of this paper is to help you to uncover some of the keys to understanding and effectively using document indexing in your content management system.

Document Indexing Exposed

Most documents that are stored in content management systems (CMS) will be indexed. Key identification information will be extracted from the documents and saved into the CMS so that the documents may be retrieved using that information later. For example scanned accounts payable invoices may be indexed in a CMS using the invoice number, invoice date, and purchase order number. Users could later key an invoice number into a search screen in the CMS and list all of the matching invoices and then click on an invoice to display it in a viewer. This type of index information is sometimes referred to as "metadata" or "template" based information. Content management systems also provide additional system indexes that may be helpful for locating documents in the future such as the date the document was scanned or imported, a document classification, typically called a document class, and the department, name, login ID, and workstation name of the user who originally captured the document. These types of indexes are captured automatically by the system as documents are added to the CMS.

In addition, many content management systems provide a content search capability so that documents may be located by searching for words contained within the documents. This type of search is helpful for documents that are more of a free form format such as letters or other electronic documents such as e-mails. Many systems provide complex content search capabilities that allow users to specify rules that are to be used to locate documents. For example a user may want to display all documents that contain a certain word but omit another. Or they may want to display documents that contain a certain word within a maximum proximity of another word in the same document.

Some Bare Facts of the Benefits and Costs of OCR

Documents that are already in an electronic format such as e-mails or spreadsheets are easily made searchable in a CMS through the use of filter technology. Filters are small programs that extract text from documents as they are checked into the CMS. They extract the text that would be helpful for locating documents later. Some systems keep track of the exact page and position location of the text within the original document while others simply extract all of the text from the document. In order to provide the ability to search for content in scanned images, optical character recognition (OCR) needs to

be performed on the documents. OCR is the process by which the scanned images or pictures of the letters contained within each document are turned into searchable text. OCR is a very processor and memory intensive operation. If all scanned documents are to be made content searchable the appropriate server or workstation resources must be dedicated to the OCR operation. Processing a single scanned page can easily take fifteen seconds on the fastest server and use a hundred percent of a single processor and hundreds of megabytes of memory. If the intensity of this operation is not taken into account, server and workstation resources can quickly be overwhelmed. If other operations are taking place on the server workstations or servers their performance may be severely degraded while OCR operations are taking place. Because of the amount of resources required to make scanned documents content searchable this cost has to be weighed against the benefits. There will be cases in which documents simply do not lend themselves to metadata type indexing and content searching is the only option. In each case the system architects should carefully weigh the OCR resource requirements.

Uncovering the Correct Balance Needed with Document Indexing

Document indexes provide an easy way to locate documents in a CMS. However there is a cost associated with the creation and maintenance of each document index. Document management architects try to strike a balance between providing enough indexes to make document retrieval easy while minimizing the cost of creating and maintaining the indexes. There are various methods for extracting the indexes from scanned documents. The most obvious involves simply displaying the scanned images from each document and then having an operator physically type in each index value. As the volume of scanned documents increases most companies will opt for more efficient methods of indexing documents. For instance, as noted previously, OCR may be used to extract indexes from scanned documents. While OCR technology is very accurate especially when processing clean typewritten documents it is difficult to determine where the indexing information is located on each document. For this reason most high volume document capture systems will involve the use of some type of template or rules-based index extraction system. With a template-based system an administrator will create a template that approximates the layout of each type of document that is to be scanned. Within the template they will define where each index field is and then assign a name and define a set of rules for that index field. Those rules will include parameters for the index information that is expected to appear in the field such as defining whether there are only numbers or mixed letters and numbers.

Database lookups may also be defined so that the index field is validated in a database. Rules-based systems operate without the use of templates but still require some degree of interaction with either the user or an administrator in regard to learning the layout of the documents. A rules-based system will perform OCR on each incoming document and then search a database of knowledge about scanned documents. If the knowledge database doesn't contain enough information to tell the system where the index fields exist in the document, the user or administrator will be asked questions about the document. Then the system will remember those answers and over time the number of questions will decrease as the system learns. There are advantages and disadvantages to both approaches. The template-based systems provide a high level of control over the indexing process and are typically much

less expensive than rule-based systems. But template-based systems require the creation of the document templates up front while rule-based systems may come complete with an existing knowledge base of common business documents such as invoices. In the end, both systems can dramatically reduce the amount of manual labor that needs to be spent to index documents and as a result reduce the cost.

There is an additional cost associated with the storage of metadata indexing information in content management systems and that is maintenance. As companies merge, shutdown, or are acquired by other companies the index information that has been previously stored for these documents may become obsolete. Users searching for invoices for Company A may need to instead search for Company B. Internal customer account numbers may change as number ranges run out. A proper document management strategy takes these changes into account and either re-indexes the existing documents or else creates new index fields so that the old and new values are not mixed together. Another strategy may involve linking the documents in the CMS to records in an ERP system so that the search capability within the CMS is not even used and documents are only located through the ERP system. The cost of a single audit may easily dwarf all efforts spent at properly planning and maintaining a document management indexing strategy!

Document indexing is a broad topic and one article does not really do it justice. However, the point is that by spending time looking under the document indexing covers you will start to understand how to weigh the benefits and costs of the various indexing tools. As a result, building and using a cost effective content management system will seem much less daunting.

By Joe Hill, President and CTO, UFC, Inc.

www.ufcinc.com

About UFC Inc.

UFC Inc is a consulting, integration and solutions firm preferred by clients in the Oil and Gas Industry for our quality, innovation and integration expertise. UFC provides data capture, enterprise content management software, support and integration services - based on a flexible architecture and common set of applications for collecting, classifying, retaining, migrating, securing and accessing information – all at the lowest cost of ownership.

Unlike vendors that deliver generalized ECM products with centralized or consolidated architectures, or support few applications and data types, UFC delivers the most comprehensive solution, specifically tailored for the customer. The distributed nature of the solution along with UFC's extensive expertise and unique approach makes it ideal for the Oil and Gas company with remote offices that have limited storage space, minimal IT infrastructure or technical support. Remote locations realize significant improvement in operational efficiencies, improved collaboration, a reduction in storage costs - without sacrificing centralized control or visibility of information. From capturing personnel information such as fuel cards and human resource forms to capturing and storing engineering drawings and correspondence, UFC provides the Oil and Gas industry the ability to reduce paper transaction costs while increasing their data processing efficiencies.

Call us today to find out how we can help your organization at (248) 447-0100 or email us at sales@ufcinc.com.

Keywords:

document capture
document capture software
web based scan
web based scanning
scan capture software
document capture solutions
capture document
process capture software
capture documents
web based data capture