The Document is the Database

by Kirk Olson and Mark Steel, Document Science, LLC.

Coopers & Lybrand reported some surprising statistics in a recent study:

- The average company incurs labor costs of $20 to file a document, $120 to locate a misfiled document and $220 to reproduce a lost document.
- 7½% of all documents get lost and 3% of the remainder are misfiled.
- Professionals spend up to 50% of their time looking for information and only 5% to 15% of their time reading it.
- There are over 4 trillion paper documents in the U.S. alone – growing at a rate of 22% per year.
- The average document gets copied 19 times before it is permanently archived or destroyed.

In addition, paper-based filing systems allow documents to reside in only one place at a time. Therefore, office personnel will often make their own copy of documents.

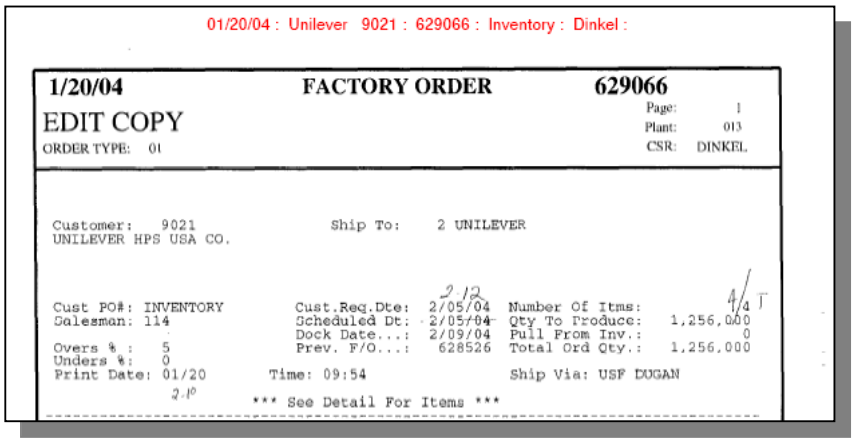
Document Imaging

Document imaging evolved because 90% of corporate information resides on paper. Document imaging is the process of converting paper documents into electronic documents that are exact replicas of their paper counterparts. They can then be indexed, stored and retrieved from the company's computer network, the Internet or individual computers. Document imaging systems are finding a warm welcome in many organizations because they can make significant improvements in operational efficiency with little organizational change. In a business climate where organizations are seeking ways to cut costs and increase productivity, document imaging systems are providing the most dramatic impact on productivity since the copy machine replaced carbon paper.

OCR is the process of converting text contained in a scanned image into text that is searchable. After a document has been OCR'd,, a "full-text search" can be executed using words and phrases known to be included within it. The OCR process is sensitive to the quality of the image, as well as the font differences within the document. As a result, OCR processing is not yet capable of achieving 100% accuracy.

Embedding a Label into the Electronic File

Many document imaging systems store documents in an independent database. This type of database can be problematic because the database index is simply linked to the location where the documents are stored. As a result, the documents cannot be transferred across multiple platforms, the system may have compatibility issues with future software upgrades, and the user may be bound to the original host imaging system.


01/20/04 : Unilever 9021 : 629066 : Inventory : Dinkel :

1/20/04 FACTORY ORDER 629066

EDIT COPY

Page: 1
Plant: 013
CSR: DINKEL

ORDER TYPE: 01

Customer: 9021 UNILEVER HPS USA CO.

Ship To: 2 UNILEVER

Cust PO#: INVENTORY
Salesman: 114

Cust.Req.Dte: 2/05/04
Scheduled Dt: 2/05/04
Dock Date...: 2/09/04
Prev. F/O...: 628526

Number Of Itms: 4
Qty To Produce: 1,256,000
Pull From Inv.: 0
Total Ord Qty.: 1,256,000

Overs %: 5
Unders %: 0
Print Date: 01/20

Time: 09:54

Ship Via: USF DUGAN

*** See Detail For Items ***

Embedded label on scanned PDF document

A better alternative is to embed labels directly in every electronic document. With this approach, everywhere the document goes, the data label travels with it. This is an enormous advantage because it eliminates the need for database maintenance and the requirement to add "meta-tags" to documents for web-enablement. An independent search engine such as dtSearch or Google can also be used very efficiently with these embedded documents. Their indexing systems simply extract the character information from the documents and generate a separate database from which searches are conducted. This allows the freedom to move documents from one directory to another without

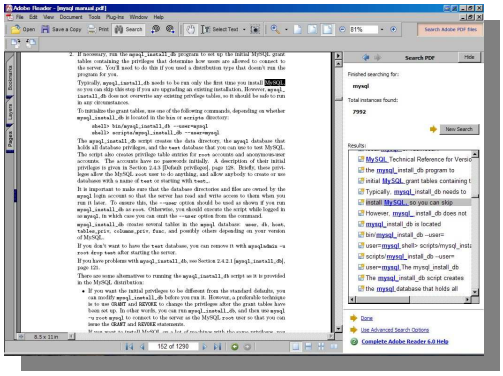
Document Science, LLC ("*DocSci*")
Turning Documents into Knowledge™
704 228th Ave. NE #603
Sammamish, WA 98074
(425) 391-3945
(425) 830-9713

concern for database indexing. By simply rebuilding the search engine index with the click of a mouse, all documents are ready for instant retrieval because the embedded label moves with the document.

PDF (Portable Document Format)

The PDF file format was originally developed by Adobe for the U.S. Federal Government to store its legacy files. Currently, the U.S. Federal Government is still the largest user of PDF technology. Most individuals have encountered the PDF format when downloading electronic tax forms from the IRS.

A PDF file is a "read only" document that cannot be altered without leaving an electronic footprint, and meets all legal requirements to be admissible in a court of law. Furthermore, the PDF file format is practical and economical in that it allows documents to be stored on a company's server. This eliminates the need for additional hardware (except for additional hard drive space) and facilitates exceptional integration into any network.



Adobe Reader

PDF format has been a de facto Internet standard. It guarantees that the image seen by the viewer is congruent across all platforms. While PDF requires a viewer, it is readily available as freeware called Adobe Reader.

PDF files have metadata (data that describes the document), such as XML tables of content and links, making images more useful to end-users. PDF files support security privileges, watermarking and signing, to protect intellectual capital.

One significant attribute of PDF format is the superior appearance of the printed copy when reproduced using a high quality printer. Image and text characteristics of PDF files tend to reproduce very well under most display and output configurations.

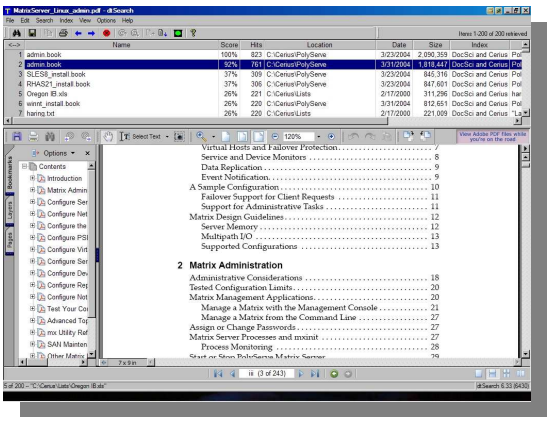
Adobe Reader with a plug-in search module can be used to search, view and print documents. Because this software is free, users avoid a "per-seat" charge for workstations connected to the network. It does not matter whether 2 users or 10,000 users are on the network; the cost of the system is the same.

OCR vs. Embedded Labels

When imaging documents, the decision to use Optical Character Recognition (OCR) versus embedded labels should revolve around the issues of data mining and document retrieval. For documents and/or information contained within those documents to be searchable, electronic documents must be indexed. OCR helps automate the process of creating searchable information. A flexible document imaging system will allow users to either label or OCR a document for indexing purposes. This is ideal because , the preferred indexing method should be determined on a document-by-document basis. Understanding the drawbacks associated with each method will help clarify which method - OCR or labeling –should be applied.

Search Results

While search engines are easy to use, the search results are often imprecise and display irrelevant information. For example, searching for “java” might return documents that describe java the programming language, java the coffee, and Java the Indonesian island. The greater the number of documents on a company’s server, the greater the number of irrelevant search results users are likely to experience.



dtSearch

Document Science, LLC (“*DocSci*”)
Turning Documents into Knowledge™
704 228th Ave. NE #603
Sammamish, WA 98074
(425) 391-3945
(425) 830-9713

Embedding labels into documents can significantly improve the efficiency of searches by returning a higher percentage of relevant documents. This is especially imortant for companies operating in industries where standardized documents, such as human resource records and invoices, need to be retrieved. For other organizations, it may be more practical to OCR the documents in order to search for a keyword or phrase contained within the text.

Document imaging systems should allow intra-document searches. This is especially useful if one is data mining or looking for information contained within a document.

One can retrieve a document and then jump directly to the specific information needed within that document. Navigation from occurrence to occurrence of the keyword(s) will start with the first "hit." Each "hit" should be visibly highlighted within the document, making the search and retrieval process much more efficient.

By embedding a label into a document and using OCR with pages that are relevant to a future search, your document database will remain a manageable size and become a document driven database.

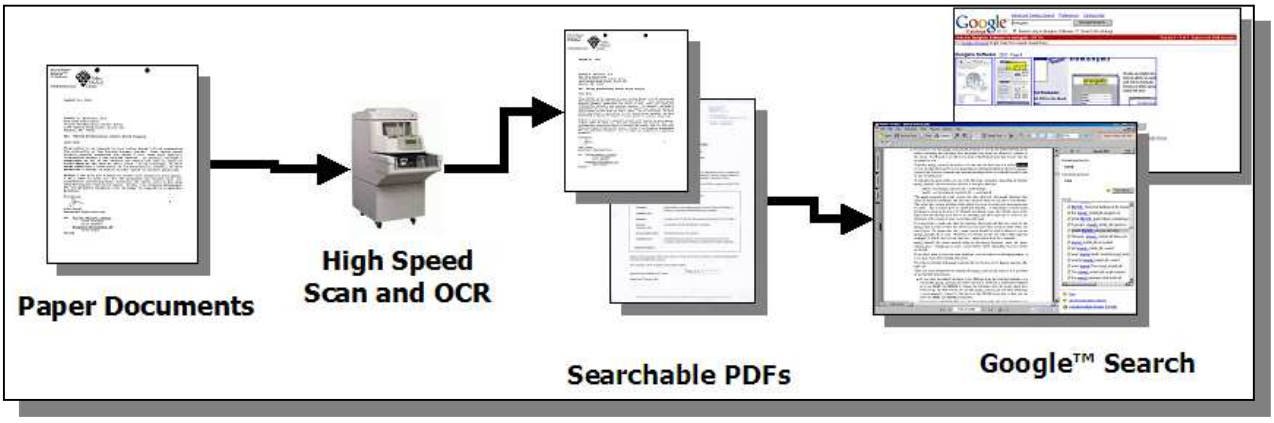
Many users find it faster and more accurate to manually enter predetermined keywords, phrases and numbers to index each document rather than to OCR each document and correct the suspect characters. The freedom to use an unlimited number of characters and numbers when labeling documents allows operators to use existing terminology with which employees are already familiar. This facilitates a quick and easy transition into using digital documents in place of paper documents throughout an organization.

Simplicity is the Foundation for a Quality Document Imaging System

There is great value in the straightforward simplicity of a document imaging system that uses the document as the database. The process starts with a paper document being fed through a high-speed scanner. Then the image is converted into an electronic PDF file. This industry-preferred file format is ideal for sharing across a company's local area network (LAN), accessing over the Internet, or archiving to a hard drive or other media. In addition to being a document conversion solution, our document imaging system also integrates with the Windows printing system, allowing users to create a PDF file directly from any application by using the "print" function. This completely eliminates the need to print documents to paper and then scan them back into the system.

With a simple document imaging system, an entry-level employee can be in full production (scanning, labeling, and indexing) after less than two hours of training. A simple systems should also require only basic computer skills for installation and management without the need for constant intervention of IT personnel.

Everyone on a network can retrieve, view and print PDF documents created by a document imaging system utilizing Acrobat Reader. Company personnel can perform these functions from their desktop PCs after less than fifteen minutes of training. Most people are already familiar with Acrobat Reader because it is standard software preloaded on most personal computers and is widely used when navigating the Internet.



Document Science, LLC ("*DocSci*")
Turning Documents into Knowledge™
704 228th Ave. NE #603
Sammamish, WA 98074
(425) 391-3945
(425) 830-9713

Electronic Document Storage and Retrieval

The ultimate objective of any document imaging system is to scan a paper document and convert it into an electronic document that can be searched, retrieved and shared with other people across a network, the Internet, or an intranet. Eventually, the documents are removed from the active system and archived. When archived, it is critical that they are in a format that can be searched and retrieved easily. Two categories of documentation directly related to the storage of electronic documents must be addressed when managing a document imaging system: current, or active documents, and archived documents.

Active Documents

Current or active information should be stored on the local server, affording the security precautions associated with a network and providing instant access to the documents. An advantage of the PDF document in this environment is that numerous users at different workstations can view a single document on a network simultaneously.

The hard drive capacity required to store current data is dependent on the number of electronic documents and the length of active time assigned to each individual file.

Archiving Electronic Documents

With time, documents become inactive and will be removed from the active server system and archived. These documents can be archived on a tape drive, a CD-ROM or on an optical disc.

Disaster Recovery Protection

When a disaster strikes, whether a fire, flood, tornado or other event, the initial damage is only the beginning of disruptions that can last weeks to years. Gartner Research recently released a chilling statistic: two out of five enterprises that experience a disaster go out of business within five years. The absence of adequate backup plans hampers recovery efforts, resulting in significantly greater recovery time - if recovery is even an option. With a document management system, the solution to document disaster recovery is simple; download a copy of all the documents onto a new directory, re-create the index, and full access to original documentation will be restored in a matter of minutes.

One of the inherent benefits of an electronic document is that a copy of it can easily be created and stored off-site. Although enterprises routinely backup their servers, the problem lies in that 90% of an organization's knowledge resides in paper. The advantage of our document imaging system is that it converts these paper documents into electronic PDF documents, which can be copied and stored off-site. The return on investment for this benefit is not measured in dollars and cents, but in peace of mind.

Summary

Using embedded labels in addition to OCR allows any business to build a database containing the full text of all their critical documents. Documents can be scanned, indexed, searched and manipulated to speed the information retrieval process. Truly, the document becomes the database. It is retrievable, printatble, portable and shareable.

Stay tuned for the next in this series – "Scanning or copying?" and "How Digital Storage can help with your Business security policies and disaster recovery plans"

The authors: Kirk Olson and Mark Steel are cofounders of Document Science, LLC. "DocSci" provides high speed outsourced scanning services, consulting and offsite hosting.

Document Science, LLC ("*DocSci*")
Turning Documents into Knowledge™
704 228th Ave. NE #603
Sammamish, WA 98074
(425) 391-3945
(425) 830-9713