What to do with all those documents?

(or how to build a "Googleable" database)

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

Your problem is too many documents and no easy way to find the information you need. Today's business processes create more paper than ever before. Paper creates problems when you can't find the information you need, when you need it. Copying, reviewing and summarizing documents is also time consuming and expensive.

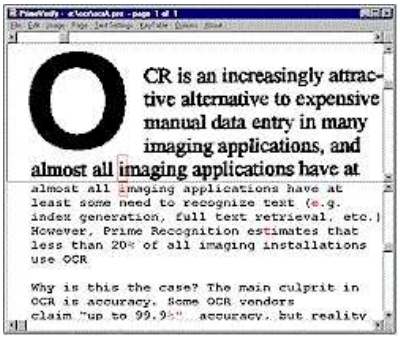
Previous attempts at document management solutions included complex systems that provided good summary data but no "ad hoc" access to unstructured information. Google™ and other new search technologies have come to the rescue.

Six technologies have emerged in the past year that have changed the game dramatically.

- Significantly better Optical Character Recognition
- Improvements in PDF (Portable Document Format) for scanned documents
- High-speed search engines
- High-speed scanners
- Dramatically lower cost on-line storage
- High-resolution LCD displays

Optical Character Recognition (OCR)

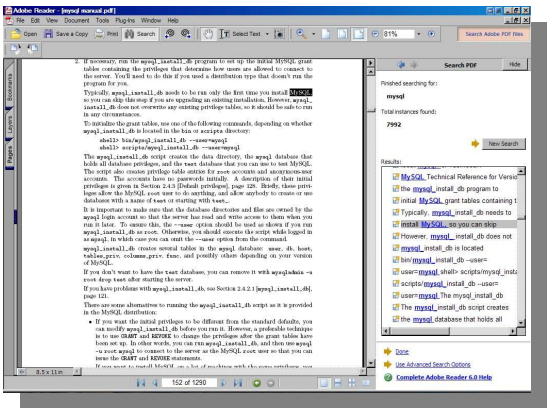
OCR software "reads" scanned images and converts them to text. Only five years ago, this technology provided an average of 90% in text recognition accuracy. This meant that a page containing 2000 characters averaged 20 errors. Today's OCR engines offer over 99% accuracy; or one or two errors per page.



©Document Science, LLC. 2004

PDF (Portable Document Format)

Probably the biggest single breakthrough in document imaging technology is PDF. In the old days (three years ago), two format options were available for scanned images: TIFF, offering a true digital image of a document or OCR, a conversion to a plain text format with no formatting. One was a good facsimile of the document, but with no text capabilities, and the other was text, which probably had little resemblance to the original document. With PDF, documents are scanned, OCR'ed and converted to digital text information, and a true "copy" of the document is preserved. Metadata (descriptive data) is also embedded in the PDF file format. This means that, every word of your document is stored, along with the location of words in the document. This is a huge breakthrough; --PDF documents meld the two old options into one format that is both a true image and totally searchable.



PDF storage also allows the user to:

- Store other metadata, including summaries.
- Apply draft, confidential and other types of digital stamping.
- Apply digital signatures.
- Append and delete PDF pages.
- Combine multiple PDF files into a single PDF file.
- Split large single PDF files into multiple PDF files.
- Convert PDF documents to HTML, text and word files.

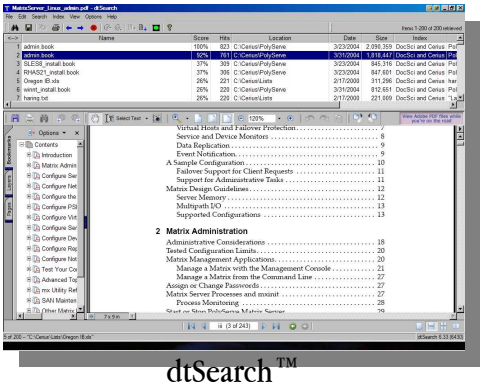
The Portable Document Format has been around for a while, however what has changed are the toolkits available manage these files. A number of free tools have emerged for PDF creation recently. For example, PrimoPDF (www.primopdf.com) and PDF995 (www.pdf995.com) are free tools that allow you to "print" to PDF format. Word™ documents, web pages

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

and any other printable documents can be converted to PDF files using these tools, laying the foundation for an unstructured information database.

High Speed Search Engines

Google™ has changed the way people find external information today. However, most businesses have failed to employ the same approach with their own internal information. Search engines that run on desktops or servers can change this. Scanned documents (or any other digital data) can now be



dtSearch™

accessed immediately using a Google™ Appliance or other search engines. These engines allow instant access to specific information within:

- Scanned Documents in PDF format.
- Word™, Excel™ and PowerPoint™ documents.
- Outlook™ emails.
- Any other text-based files.

Some of these search engines are free. Most are low cost. The current king of searching, Google™, is now available as a network attached appliance. By connecting the Google Appliance to a network a business can "googlize" its entire document repository.

A rudimentary search engine is built into Adobe Reader™ (aka Acrobat™). This software allows searches within a document or a group of documents within a folder. Searching is relatively slow compared to Google™, but still effective when top speed is not crucial. Another search option available to Windows™ users is dtSearch™. The dtSearch software is a $200 package that builds an index of words from any textual file. Searches with dtSearch are instant and the "hits" are displayed immediately.

Comparison of Search Tools

Search Engine	URL	List Price	Features
Built in Adobe Reader™ search by Onix	www.adobe.com	Free	Fast serial search, built in, all platforms
DocSearch, Java based search	www.brownsite.net/docsearch.htm	Free	Builds external index, all platforms
dtSearch™	www.dtsearch.com/	Single, $199; Network, 5 user $800	Builds external index, works with all kinds of documents, passes search criteria to Reader, Windows only
ARTS PDF Search	www.artspdf.com/arts_pdf_search.asp	$1,300 unlimited users	Builds external index, works with all kinds of documents, Web based interface,. Windows servers only.
Google ™ Appliance	www.google.com	$32,000+	Builds external index, works with all kinds of documents, extremely fast

High Speed Scanners

The current generation of high-end scanners includes models from Fujitsu, Canon, HP, Bowe Bell and Howell, Kodak, and others. Speeds up to 480 images per minute are possible with the latest production document scanners. However, scanning hardware accounts for only about 20% of the cost of imaging. Manual processes, including document preparation, indexing and returning documents to their original condition, make up the rest.

Fujitsu 4099D

©Document Science, LLC. 2004

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

Low cost on-line storage

Storage costs have dropped below 50¢ per Gigabyte (1,000,000,000 bytes). A 200 Gigabyte disk can hold over 1 million scanned documents.

High resolution LCD displays

LCD, or flat panel, display technology allows large area screens to sit on the desktop. Resolution up to 2560 by 1600 pixels allows excellent quality, non-flickering display of many documents at once.

Summary

The synergy of new and improved technologies has created the ability for the average business to build a document database containing the full text of all their business documents. Documents can be scanned, searched and manipulated to speed the information retrieval process. Digital documents can also be shared, annotated, merged, marked-up, printed and emailed far more efficiently and economically than the hard copy originals.

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. 2004

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