White Paper – What You Need to Know About Document Conversion

2-4-14 V1.1

Introduction

The term "paperless" was first used in a June, 1975 BusinessWeek article titled "The Office Of The Future". But over time, rather than eliminating paper we actually have more of it around today than we did before! Many people think the proliferation of mobile devices will help drive us to a paperless world but that doesn't appear to be the case either. Printing from mobile devices is one of the most requested applications corporate users ask for.

In a recent survey by the industry organization AIIM, respondents felt that driving paper out of a business process would improve the speed of response by a factor of 4. Those with more experience in paper free processes report even greater speed up of 4.6x (AIIM). So ask yourself, "Why do we still rely on paper when it's so inefficient?"

The Cost of Paper

While it's necessary to store some physical files if you walk around any office, large or small, you'll see rows of file cabinets along the wall or special 'filing rooms' that contain office documents. Have you ever thought about the cost of maintaining these file cabinets for weeks, months or years? The cost can be tremendous! An average file cabinet takes up 7.6 square foot of floor space. According to commercial realtor CB Richard Ellis, Chicago's annual total occupancy cost per square foot was $49.15 in 2012[1]. That means one file cabinet costs over $370 per year just for the floor space it consumes. That doesn't factor in the cost of the cabinet, filing supplies, and the labor costs of the clerk(s) who manage these files. Multiple studies have shown when you factor in all the costs, it's easily a few thousand dollars per year per file cabinet.

What if all your paper documents aren't simple office documents? What if you have books, historical documents, or newspapers? According to the Harris County, Texas District Clerk[2], the cost for preserving invaluable historical documents range from $10 for a historical document file to as much as $2,500 for a civil index book.

Over the years many companies turned to microfiche or microfilm as long term alternatives for archiving. While it's recognized as a stable long term medium, special equipment is usually needed to both read and duplicate the images so they are not practical for every situation. We'll talk more about this later in the paper.

---

[1] Prime Office Occupancy Costs study (CB Richard Ellis, Dec. 2012)

[2] http://www.hcdistrictclerk.com/common/historicaldocument/historicaldocuments.aspx

To try to offset these costs, some companies outsource their document storage. While this can reduce or eliminate the cost of file cabinets and on-going maintenance of files, it still requires internal staff to collect and organize the files for movement to the off-site vendor. And using an offsite vendor often contains hidden costs. In a 2011 article, Inside Council magazine reviewed a quote one of their clients received for storing documents offsite. What they found was a standard legal or letter carton is 1.2 cubic feet, and at a monthly storage cost of 19 cents per cubic foot each month, storing 1,000 cartons for three years would cost $8,208. But there were a number of 'extra' costs that significantly drove up the costs.

- Initially moving the boxes to storage ($3,564)
- Move labor ($411)
- Fuel surcharges (about $200)
- Receiving and entering the boxes ($2,052)
- Monthly administrative fees ($2,260)

Even worse, when it was time to get rid of the boxes, there were additional charges for destruction of the documents ($2,484). If you wanted your documents returned to you there was a fee for permanent withdrawal of the records ($3,099).

Storing and managing paper isn't without its costs or risk.

Risks with Paper

One of the major challenges with paper is it doesn't lend itself well to collaboration. Only one person at a time can possess a paper document. With today's extended workforce, and the focus around sharing of information, paper becomes a bottleneck in any business process. Documents must be passed from person to person, creating opportunities for mis-placed and lost documents.

There have been a number of studies that looked at the cost of time wasted searching for documents, and the cost of recreating lost or missing information. In it handbook, "The Paper Free Process Revolution Handbook", AIIM quotes a study by PricewaterhouseCoopers indicating that on average, 7.5% of all documents are lost and 3.5% of the remaining documents will be misfiled. In short, you will not find 11% or more of your paper-based information. In a widely quoted research report, IDC estimated a 1,000-employee organization will waste between $2.5 and $3.5 million annually simply searching for nonexistent information, failing to find existing information or recreating information that can't be found.

If you're in a heavily regulated environment like Financial Services, Government or Healthcare are these statistics you can accept? There have been numerous public examples of companies losing documents, mis-placing them, and even worse exposing personal private data. Copies get made of paper documents and uncontrolled copies begin to float around the organization. Paper documents are left on fax machines or desktops for anyone to pick up and view. Boxes of records being shipped to offsite storage locations have been lost or mis-placed. Plus paper does not leave an effective audit trail. If your company was audited, or served with a discovery request how confident are you that you could deliver

the information requested in a timely and accurate manner. What's the cost to your business if you don't?

Moving to a Digital World

These risks and inefficiencies are why many companies are moving to digital storage of documents and records. By storing records digitally, information can be easily located, shared, and collaborated on. You can control how and where information is shared, internally and externally. Audit trails can be captured that ensure corporate policies and guidelines are followed in the use of information. But before you begin simply scanning every document in site, there is planning that needs to take place.

The first step is to analyze what type of documents you have and which ones will be converted to digital form. If everything was a simple 8.5" x 11" office document that was created in MS Word and printed on high quality paper, digital conversion would be simple and almost 100% accurate. But that's never the case. Ask yourself the following questions:

1. Do you do business in other countries?
2. Do you have lots of legal documents?
3. Are your documents multi-lingual?
4. Do you have large format documents like engineering drawings?
5. What is the quality of your documents? Faxes, copies of copies?

If you do business internationally chances are you'll have to deal with A4[3] documents and your documents could be in multiple languages. During scanning you'll have to carefully sort documents so the right scan settings are used for the proper document types otherwise your results will be very inconsistent. For multi-lingual documents you'll have to have operators who can interpret the documents when sorting them for scanning and indexing. If performing optical character recognition (OCR) to extract the document text or automate metadata extraction you'll need software that can understand multi-lingual documents. Language styles differ, the simplest example being U.S. English and British English: colour or color, analyze or analyse?

When looking at metadata extraction, consider the layout of your documents. Documents typically fall into three styles: structured, semi-structured, and unstructured.

- Structured – this includes forms and many large format documents. They have a consistent layout in their design and the information for extraction always appears in the same place on the document. This allows for a high level of automation to efficiently capture and extract the metadata with minimal manual intervention.

- Semi-structured – documents like invoices and purchase orders have consistent layouts, but are not as structured as forms. Terms like PO Number, Account Number, Total, etc. may appear in

[3] Visit http://www.papersizes.org for more information on international paper sizes

various places on the document and be abbreviated differently. So while automated tools may be used with a fairly high rate of accuracy, manual validation of extracted metadata is often still required.

- Unstructured - Contracts, correspondence, legal documents, books and handwritten pages require a more manual process to extract accurate metadata. While you may be looking for certain keywords in the document, they can appear anywhere in the text and be formatted or referenced in different ways. While automated tools can help in the process by identifying potential keywords, manual work is required to identify and validate the metadata.

Using automated tools can speed up the process of metadata extraction but it's likely they won't be able to extract 100% of the data you require. Often one metadata field may need to be populated with information found in multiple locations throughout the document. This is common in title and mortgage documents, requiring manual review and analysis of the data. Lastly think about handwritten documents. Handwriting styles vary significantly from person to person. While you can easily scan a handwritten document, transcription may be required to get an accurate electronic copy of that document.

Now think about other types of information that may need scanning. Do you have large format documents like engineering drawings you need to digitize? These will require different scanners than document scanners. File sizes will be larger and special viewing software may be required due to the size and format of the documents.

What about non-office documents? Do you have books, newspapers, or historical documents you want to digitize? [Company] has done a number of conversion projects where we have digitized newspapers, historical documents, and books for our clients and made them available online. But these documents require a different approach than digitizing office documents. Depending on the age and importance of the documents special handling and training may be required to prevent damage to the original documents. Each document may require unique adjustment of the scan settings to get the best results and obtain a usable image. Again, special scanners may be required for books or newspapers.

In addition to dealing with the variation in sizes and formats you also need to factor in the quality of your documents. Are all of your documents first generation documents? Are they faxes or copies of copies? Is the document skewed at an angle? Is there background noise from poor copying? Are the documents old and aged? All of this will affect the end use of the document and has to be taken into consideration when planning a digitization project. We'll discuss this further later in the paper.

Microfilm and Microfiche

An early method of moving away from paper was microfilm and microfiche. Microfilm was recorded on small reels of film while microfiche is a flat sheet. These are commonly used in archives and libraries to provide access to specialized information, limit handling of rare or deteriorating documents and to save space. While more efficient than paper, there are challenges with these formats. Because of their size,

neither are readable by the naked eye and require specialized readers that magnify the images on the film. The readers are bulky and are far less common than personal computers, so it requires training of users. Plus, making a paper copy of a document stored this way also requires using a special printer integrated into the machine. Because the film is simply a picture of the document it's not searchable in any form. Manual or electronic indices have to be kept and maintained that list which film contains what information. So the process of retrieving information looks something like this:

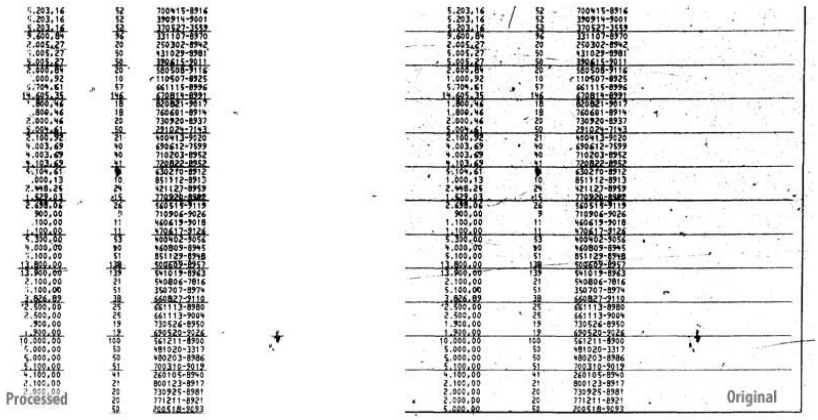
1. Go to the index and search for the information you're looking for. The index could be electronic or physical.
2. Write down the information on what specific microfilm reel or microfiche image you want. Do this for each document.
3. Submit the request to the librarian or archivist. Wait for them to go retrieve the film.
4. Sit at the reader and load the film.
5. Navigate to the right image on the film and read the information you're looking for.

Not very efficient is it? And if you need a reference copy of the information you know what you have to do? That's right, print a paper copy or take very good notes.

Image Quality and Format

Earlier we briefly discussed document quality. Let's look at it in more detail now. Every time a document is copied or faxed the quality and contrast of the document is reduced. In addition other factors can be introduced such as skewing or background noise. All of these factors directly affect the quality of the scanned image and make the scan operator's job more difficult. The poorer the image quality, the more difficult it is to get good results when extracting metadata or performing OCR on the documents.

We've all seen copies of a fax or 3rd or 4th generation copies of a document. They look nothing like the original and often are skewed at an angle, and have 'noise' in the documents. Skewing is what happens when a document is faxed or copied at a slight angle. Each subsequent copy or fax can exacerbate the problem. When referring to documents or images 'noise' is typically viewed as black bands along the edge of the documents from poor copying, dots or black spots scattered about the document, and even missing information.



This causes problems later on in the digitizing process. If you want to automate the extraction of metadata from the documents or perform OCR on the documents, they must first be digitally cleaned up. This means both de-skewing and de-speckling the images at scan time or afterwards. While the software tools available today are very good at these processes, manual inspection and QA may be required for extreme cases.

The reason this has to be done is that OCR software is highly dependent on image quality. If the contrast of the document is poor, if it's skewed, if there is background 'noise' in the document, OCR quality will suffer. In poor contrast documents the OCR software may not be able to detect the edges of characters or where words start and stop. If a document is skewed at an angle, the OCR software may not recognize letters or words. For example is an l really an l or is it a /? Noise may be interpreted as characters or cause the software to mis-identify a character. An O could easily become a Q for example, an l an i. This increases the cost and time required to get accurate data extraction from your images.

The last thing to consider when digitizing documents is the format you want to store the documents in long term. This can have a direct impact on a number of factors including:

- File size
- Quality
- Viewing software
- Search

The most common form for storing documents today is portable document format or PDF. Developed by Adobe[4], PDF is an open standard for electronic document exchange and is maintained by the International Standards Organization (ISO). Free viewers are available from a variety of sources across multiple platforms, including mobile devices. For long term archiving many companies have adopted the PDF/A standard, a variation of PDF designed specifically for this purpose.

An alternative is an image format such as tagged image file format or TIFF[5]. This is a common format for image files used in the graphics and publishing industries that offers a number of features. Originally developed in the mid-1980s, it's a mature format that has seen few changes in recent years. Many free viewers are available for TIFF as well. Unlike PDF, TIFF is not an open format. Adobe holds the copyright for TIFF.

In addition to the image format, you need to determine the best format to capture the image metadata. Many different formats exist, and the system you'll be importing the images and metadata into will have a direct impact on your decision. The text plus images may need to be in:

- HTML – standard Hypertext Markup Language (HTML) for use on the web
- XML – Extensible Markup Language for use in a variety of applications

[4] http://www.adobe.com/products/acrobat/adobepdf.html
[5] http://en.wikipedia.org/wiki/Tagged_Image_File_Format

- EAD – Encoded Archival Description, an international standard maintained by the Library of Congress in partnership with the Society of American Archivists.
- TEI – Text Encoding Initiative, an international and interdisciplinary standard widely used by libraries, museums, and publishers for on-line research and teaching.
- ContentDM – specialized format used with the ContentDM Digital Collection Management Software
- Custom formats – we can deliver images and text to you in custom formats such as CSV, tab-delimited or other delimiters.

[Company] experts can help you determine the best format for long term storage and archival of your image files and metadata.

Planning a File Conversion Project

When planning a document conversion project you have two choices: do it in-house or outsource the work. On the surface many firms look at doing the work themselves, but there are a number of risks to this approach, especially if you don't have document imaging expertise in house.

One of the primary barriers to internal document conversion is the overall cost of the project. Many companies under-estimate the time and cost of a large conversion project. Even if you're currently scanning documents internally, ramping up for a large backfile conversion project requires a significant investment in staff, hardware and software resources. Additional hardware resources may be required to handle the volume of documents being scanned. As mentioned above, non-standard office documents may require special handling and hardware. Software licensing costs may increase for your scanning software due to increased volume or users. Additional scan, quality assurance and keystroke operators must be hired and trained to handle the increased volume.

The second thing to consider is the quality of the document conversion. [Company] experts have years of experience in document capture, data extraction and our quality assurance processes have been developed based on best practices. How deep is your expertise in software tools like automated data extraction and OCR? A newly trained staff using unfamiliar software will surely make mistakes in document scanning leading to poor quality images and in data extraction leading to poor meta-data quality.

This leads to the third issue, project delays. Can you meet the production volume you expected to? If quality isn't at the level you expected, what's the impact on the project timeline? If meta-data quality isn't at the levels you require, what's the impact on downstream systems relying on that data? [Company] guarantees the quality of the images and date we deliver.

Benefits of Document Conversion

The benefits of moving away from paper based processes and records has been proven over time. At the top of this paper we mentioned in a recent AIIM survey respondents feel that driving paper out of a business process would improve the speed of response by a factor of 4. Those with more experience in paper free processes report even greater speed up of 4.6x. Studies by Aberdeen Group[6] have shown that automating paper based processes in accounts payables can reduce time and cost significantly, by a factor of 5 or more. Not only can you respond to your customers in a more timely fashion, you can do so at a significantly lower cost.

In addition to reduced cost, moving away from paper reduces your compliance risks. Digital information is less prone to mis-filing and loss, security can be more readily enforces and electronic audit trails are easily captured and stored. For heavily regulated industries like healthcare, government and financial services this factor alone can justify the cost of document conversion. The risk of non-compliance is just too great, financially due to fines and penalties and in negative publicity, impacting corporate image.

If you're still managing large collections of paper in file cabinets or off-site locations, document conversion is certainly something to consider. A proven ROI, lowered risk, improved employee productivity and improved customer satisfaction are all direct results of paperless processes. While it's probably not realistic for most organizations to be totally paperless, the benefits of eliminating paper where possible directly contribute to your bottom line.

About [Company]

Based in Chicago, Illinois, [Company] can assist you with your document conversion project. What sets us apart from other bulk document scanning and conversion companies is our focus on quality, scalability, expertise and customer service. Our experts help you analyze the scope of your project, define the requirements based on our experience and industry best practices, and deliver the highest quality regardless of the size of the project.

With a large staff of experienced Content Editors, [Company] has one of the largest resource bases in the industry. [Company] retains professionals with college degrees and experience in the process of data extraction. We deploy world-class quality control procedures to deliver the highest quality content. Most importantly, as a Services Organization, we take pride in our professionalism and focus on customer service. Call us today and get started on the path to a digital world.

[6] Invoicing and Workflow, Integrating Process Automation to Enhance Operational Performance – Aberdeen Group, May 2011