RHEINNER GUIDE

Configuration Criteria for Document Imaging Systems

Document Management, Imaging, and Workflow Technology Guide Series

This guide has been sponsored by

Panasonic
Computer Peripheral Company

Table of Contents

The Rheinner Group

The Rheinner Group is a leading research, consulting and education firm in the document imaging, management and workflow industry. Its Certified Document Imaging Architect (CDIA) Education Program, which covers many of the same issues addressed by The Rheinner Group's Technology Guides, is the most popular training program in the imaging industry.

For more information on The Rheinner Group, CDIA course schedules, or to obtain help designing and implementing document imaging and workflow systems, please call 617-741-8100. Also, please visit our web site, The Rheinner Café, at www.rheinner.com.

The Rheinner Group is a proud sponsor of the Certified Document Imaging Architect (CDIA) certification program.

Introduction

With the majority of information in organizations still on paper, widespread and concurrent access to that information remains impossible. By putting that information "on-line", document imaging systems allow critical business information to be shared and utilized by many people in many places. Why, then, does so much information remain on paper?

One reason is the uncertainty surrounding the design and implementation of a document imaging system. While users generally believe that converting paper to digital information is desirable, they don't know how or where to get started. And, after hearing about a few disastrous document imaging implementations, they don't know if they want to get started at all. Many of these failed systems, however, did not result from poor integration or "buggy" software; instead these systems were not properly designed and configured. With proper planning and an accurate gathering of the necessary data, designing and configuring a successful and effective document imaging solution can be a straightforward process.

This Rheinner Group Technology Guide is designed to assist you in that endeavor by identifying the information that must be gathered, outlining how to analyze this data in the context of a document imaging system, and providing equations and guidelines to facilitate the system configuration effort.

Configuring a Document Imaging System

All document imaging implementations originate with the business benefit, not the technology. This means that an analysis of the paper control problem should precede any consideration of the technical issues. The paper control problem should then drive the technical specification and integration of the solution.

There are four distinct phases to any document imaging system configuration effort:

- Discovery
- Analysis
- Design
- Implementation

In the first, and most important, discovery phase, the work process is studied in detail to determine how work flows through a department or organization. This involves understanding the tasks that users perform and the documents that trigger these procedures. Substantial data must be gathered about the documents themselves, such as how many pages are received, at what frequency and in what volumes. And the physical quality of these documents must be evaluated as well, including the condition, color, weight, finish, size and legibility. The basic objectives of the Discovery phase are to identify what documents need to be converted to images, and how the unique properties of the documents and the process will affect system configuration.

During the analysis stage, this data is evaluated to determine its impact on the capture, storage, display, management, output and communications subsystems. Based on this analysis, the subsystems can then be designed and specified. In the final stage, Implementation, all subsystems must be integrated, tested, refined and finally deployed. Building and implementing a system is a complete subject unto itself, and it is not possible to do the subject justice in this guide. That said, building a successful system will be a much easier job if each of the previous four stages are effectively completed.

Document Imaging Defined

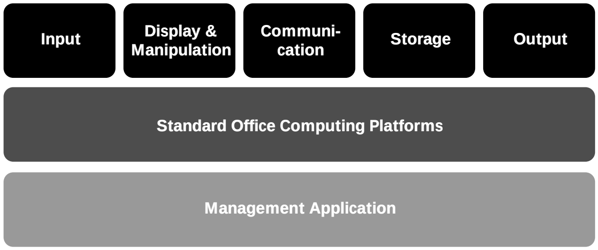
Document imaging is the electronic capture, storage, management, communication and retrieval of documents that have been converted from paper to digitized form. The management and processing of this digitized image information requires the integration of highly-specialized components with standard computing systems.

Every computing system consists of essentially five subsystems: input, display, communications, storage and output, each of which is integrated with the central processing unit (CPU) and especially configured to manage the type of data being processed. In the data processing environment, the keyboard is the input device and it is directly supported by both the computing device and the computer operating system. Each of the other subsystems is similarly optimized to manage character and graphic data, supported by the computing platform, the operating system or some other intervening layer, such as the network operating system.

But image data is very different from other computer data. Most importantly, images of documents are simply a collection of dots that form legible characters to the human eye, but are not “usable” by computer systems. This means that one cannot search for the occurrence of a particular word or phrase in an image document; one cannot spell check it; one cannot edit the document—insert, delete or rearrange the elements. The second most significant difference is that image file sizes are much larger than other computer data files. For example, a word processing document that contains a letter might comprise a file size of two or three kilobytes; the scanned version of that letter might represent an entire megabyte.

Since the unique requirements of image data are not directly supported by the existing computing infrastructure, a comprehensive array of specialized hardware and software optimized to manage image, rather than character, data must be integrated with the computing system.

Document Imaging Component Model



The Imaging Application

The imaging application, which can also be referred to as the management application, is the program the user actually sees. The management application relies on the imaging platform, an underlying foundation, to process the document image and utilize the functionality provided by the various imaging subsystems. Some imaging platforms already are management applications; others require the integration of various tools and programs to create a user application.

Document Management and Imaging

Document management platforms are not imaging platforms. The primary advantage of document management software is its ability to augment the file directory management of the computer operating system by providing a suite of extended services, including improved file security, revision control, extended file names, enhanced file descriptions, and user access privileges. Through these features, document management users gain more control over computer-generated data files, especially word processing files.

Imaging platforms, conversely, are designed to capture, annotate and manage an entirely new data type, the image file. Images require specialized viewing software, compression and decompression software, drivers for specialized subsystems, as well as support for a range of storage facilities generally found in imaging environments only.

While document management and document imaging software function in a similar manner, in that they are both file managers, imaging software comes equipped with the necessary complement of device drivers and image subsystem support required for the specialized nature of image data files. Additionally, some document imaging software applications are designed to handle millions of image files, a situation which would stretch the capabilities of even the most robust document management package.

It is possible to add image viewing software to document management applications to view, annotate and store images, just as it is possible to store names and addresses within a spreadsheet application—some people do. But the spreadsheet program was not designed to take the place of a database manager. Similarly, while you can manage document images with a document manager, document managers were not designed to take the place of an image management system.

Workflow and Imaging

Workflow is a technology tool that enables business process management and automation, by assembling individual work items into a single "work package", coordinating the assignment and distribution of these packages to the individuals responsible for their processing, and managing the completion of this processing. While the terms workflow and business process reengineering (BPR) are frequently used interchangeably or together in the same sentence, they are not the same thing. BPR is concerned with solving organizational problems through a combination of human, technology and operational methods, one of which is using workflow to automate business processes.

Workflow software and imaging systems can exist independently from each other, but they are also quite frequently implemented together. Because many paper-driven processes are group-oriented, the fulfillment of that process can benefit from the routing, queue management and work tracking provided by workflow software.

Defining and Designing the Application

All successful imaging applications start with the paper trail. Remember the maxim: "Garbage in, garbage out". Possessing a solid understanding of the papers' physical characteristics, locations, arrival rates, and processing trails has more to do with the success of a system than your choice of any technical subsystem or component. An inadequate understanding of the existing paper process, including the problems with that process, generally results in the inadequate design and implementation of the technical solution.

In data processing, it is common to configure the technical solution and then organize the data to fit that solution. In image processing, the exact opposite is true. It is not feasible to set up a system, scan the documents, and then organize them in the system. Virtually every one of the five subsystems (input, storage, communications, display and output), not to mention the core imaging application, depends heavily on the paper process and the paper problem. How quickly do you need to get document images into the system? What are the document imaging indexing variables necessary to guarantee that users can quickly and accurately retrieve the document image they need? Which storage devices and storage management techniques have to be employed to guarantee that workers don't spend valuable time waiting for images to show up on their workstation? The answers to all these questions, as well as the design parameters for all system components, are driven by the document itself. What's required is a fundamental and complete understanding of the nature and role of the documents, including their appearance and movement through a workgroup, department or organization; and an intensive evaluation of the work that is performed on those documents, and any decisions that are based on the documents.

Discovery: Obtaining the Required Document Data

Discovery, the first and most important step in the document imaging implementation process, determines all the core functional parameters that eventually dictate the technical solution. The objective of the discovery process is to identify the characteristics of the documents that will be imaged, including their physical properties, as well as the current usage patterns and processing cycles that are associated with these documents.

Document Quality

Among the most fundamental issues is the physical quality of the documents. Bear in mind that the scanner must physically move each document through a paper-handling facility without jamming. Damaged, torn, or folded paper may need to be repaired or copied before it can be fed through the scanner. And any staples, rivets or paper clips on the documents must be removed prior to scanning.

Additionally, in order to create a readable image, the scanner must be able to differentiate light from dark on the document page. That means that there must be adequate contrast between the foreground and the background of a document page. If not, special adjustments, which slow down the input process, are required. Be sure to note any other special characteristics, such as odd paper sizes, or colored or heavy paper stock. These characteristics require specialized document handling and/or image enhancement, both of which impact your choice of scanner equipment as well as the time period required to make documents available to the imaging system, an interval referred to as the document capture cycle.

Document Types and Counts

Documents should be grouped and categorized by type: invoices with invoices, contracts with contracts, etc. Alternatively, there may be several types of documents that all belong together as part of the same "case" or "folder". If they all already in folders, some estimation of folder completeness must be made. Other non-paper document types, such as jacketed film, roll film, microfilm and microfiche, must also be inventoried. At the later design stage, you'll need to decide whether or not to include these in the system.

While most business documents are single-sided, there are some important business documents that are double-sided, for example, double-sided loan applications or claims forms. While each document is only one piece of paper, it represents two image pages. This has implications for scanning and storage. Double-sided pages can be scanned in one pass on a duplex scanner, or in two passes on a simplex scanner. Scanning a two-sided page in one pass is generally more productive. However, duplex scanners are generally more expensive.

Not all business documents are black and white. Some are multi-color, such as brochures and marketing materials. Some have colors on the pages, such as invoices, bills of lading and business forms. Color pages affect several image subsystems. If the image needs to be stored and viewed in color, a color scanner and color monitor will be required. In addition, color image files are much larger than black and white image files. Therefore, they will take longer to transmit over the network and will require more storage space. If they need to be printed in color, a color printer is required.

All documents should be physically inventoried and tallied. While it is not necessary to individually count every document, you should sample a sufficient number so that quantity estimates are fairly accurate. Don't configure a system based on unaudited, internal estimates. This can result in severe under- or over-estimates, which can have dramatic implications on system requirements and performance. A general rule of thumb is to estimate 2,000 document pages per file drawer, and about 2,000 document pages per cubic foot of file storage.

Frequency of Arrival

When, how often and where documents arrive is another basic, but frequently overlooked, parameter for imaging system design. While annual document volumes are often used as a benchmark, it is very dangerous to apply these statistics without knowledge of document arrival peaks and valleys. For example, it is erroneous to conclude that out of the 1.2 million documents a company receives each year, 32, 876 arrive like clock-work each day. It may be the case that 60% of those 1.2 million documents arrive during the last two weeks of the year. The document capture process and document capture subsystem requirements will be dramatically different for those two instances. When looking at irregular document arrival patterns, it is most critical to identify and understand the arrival peaks.

Document Usage

Obviously, the physical path a document travels is of great significance. The kind of work performed on the document at each destination is of equal importance. The discovery log should record all physical locations to which the document is routed, what task or activity takes place at each stop, how the document gets from one place to another, and the time period for each step, as well as for the entire document cycle. Make sure you understand what happens at each step: Is a document inspected, referenced, annotated or foldered? Is a document put to the side until some other document arrives or action occurs? What other documents or data sources does a user rely on to complete the work associated with the document?

Analyzing System and Subsystem Requirements

With the "raw data" of document statistics collected, it's now time to analyze these findings within the context of an imaging solution. Since the eventual solution will only be as good as the sum of its parts, it is imperative to get the required level of performance from each subsystem. The best way to do this is one subsystem at a time. The last and final step, of course, is to get all these subsystem to work together as a system, which is heavily dependent on the configuration of each sub-system as well as on the choice of an imaging platform.

The analysis should initially focus on developing a digital replica of the current manual process. Each process that is now done by hand should be replaced with a digital counterpart. As the analysis proceeds, it may become apparent that certain steps are no longer necessary once the document is imaged. This may be a direct result of imaging the document, for example eliminating the need to photocopy documents, or an indirect result of linking the imaging system with a related data processing application.

To maintain focus, it is best to postpone this process redesign until the analysis stage is complete. A successful analysis requires the consideration of numerous, complex variables. The introduction of process improvements and shortcuts can raise complexities that interfere with the analysis and functional design processes. Instead, analysis and design should be iterative processes, repeated several times to uncover erroneous assumptions, misinformation, or business process inefficiencies. Once the overall system functionality is outlined and understood, this iterative approach leaves plenty of time to eliminate process steps and system elements.

Analyzing Capture Requirements

The first step is to analyze document capture requirements. The total annual document volume, peak processing volumes, and the time interval within which document images must become available to the user determine the capacity requirements for the capture subsystem. The shorter the time frame within which images need to be made available for retrieval, the higher the overall performance required by the capture subsystem. Keep in mind that several factors impact the ability of the capture subsystem to perform at a certain speed.

- *Document Preparation:* The size, quality and condition of documents all have a direct correlation to how quickly you can scan documents and make them available for the user. A separate scanner to handle exceptions may be required if there is a large number of odd-sized or damaged documents.

- *Actual Scanner Speed:* The “rated scanner speed” provided by the manufacturer is a relative, not absolute, indicator of performance that is typically achieved only under ideal conditions.

Documents are different, and therefore the speed with which they are scanned varies. Actual scanner throughput will be 25 to 40% lower than the rated scanner speed, depending on the types of documents. The only way to determine the actual scanner speed you will achieve is to scan sample batches of the documents that will be scanned.

Problem documents that have to be hand fed will take 30% longer, on average, to scan then documents that can be automatically fed. The document type and quality also dictate the need for other scanning requirements that impact capture speed. For example, some documents will have to be scanned in grayscale, which is slower than bitonal (just black and white) scanning.

Many documents can be scanned in at a resolution of 200 dpi; others will require scan resolutions of 300 dpi. The higher the scanning resolution, the slower the scanning process. Depending on the scanner, scanning at 200 dpi can be twice as fast as scanning at 300 dpi. There are, however, some techniques to speed 300 dpi scans. For example, by feeding pages in landscape (sideways) and automatically rotating them in the computer after they are scanned, you can often achieve the same speed at 300 as at 200 dpi. Since these work-arounds affect both scanner specifications and document capture cycles, they need to be identified up front.

If documents are located in different locations, it may be efficient to set up several self-contained capture stations. These capture stations should be limited to the areas where there is a critical mass of paper. While it is easier to move a scanner than paper, the process of capture involves much more than the simple act of scanning, and too many distributed scanning operations can make the input process difficult to manage and control.

- *Quality Control:* Depending on the document imaging application, the quality of the source document, and the image quality achieved with the scanning system, an additional quality control (QC) step may be required to verify the readability of the documents and to perform any image enhancement aimed at improving the readability of the document. In some cases, source document quality may be so poor that documents have to be rescanned with different scanning parameters and various image enhancement techniques before the document can be committed to the system. Some routine image enhancement techniques, such as deskewing to straighten out crooked pages; rotation to correct pages that were inadvertently scanned upside down; and despeckling to remove odd specks of dust and dirt; may need to be performed before the image is sent to the system. In other cases, image enhancement can be performed on an as-needed basis by the user after retrieving the document image.
- *Indexing:* A document index is a descriptive reference used to locate, reference and retrieve document images. Much like a library uses a card catalog to provide a descriptive index to locate all the books kept in the library, the imaging system uses an index to describe and locate an image. Needless to say, an efficient and accurate index is critical. Without it, images can easily be lost, misfiled and rendered unretrievable.

  Information gathered during the discovery phase, such as the work performed on a document and the information used to process the document, should be used to create an efficient and reliable index for the document. The simpler the index, the more accurate it generally is. Because of the critical role played by the index, there is a temptation to over-index an image with a complex array of keywords, phrases and numerous field identifiers. Not only is this time consuming, but it is also prone to data entry errors, defeating its original intent and purpose. A moderate amount of index information can be automatically provided as a byproduct of the scanning process. For example, through batch identification sheets, the system can be set up to automatically determine and record the type of document being scanned, as well as the date and time.
- *Data Extraction:* The process of extracting the data resident in an image happens at scan time. Frequently a subset of the indexing process, data extraction may also be a step unto itself, particularly in a forms-processing environment.

  There are basically two kinds of data extraction, manual and automatic. Manual extraction requires that an operator (usually the index operator) view the image on screen and "extract" the reference or other information from it, in order to key it into an index field or into some other kind of database.

  Automatic extraction utilizes an optical character recognition (OCR) engine or forms processing software to automatically extract information from the image and place it in either an index field or a database repository without any human intervention.
- *Information Release:* When the image reaches the information release stage, it has completed the document capture cycle and can be made available to the rest of the imaging system. Data extracted from the image may be released at this point as well.

Determining the Net Effective Capture Throughput

The final step of the capture analysis phase is to determine the net effective capture throughput, which is the number of images that have completed the capture cycle and are released to the imaging system per minute. This figure is determined by considering each of the steps in the capture cycle individually, as well as the actual scanner throughput. For example, if document preparation renders 30 pages per minute (ppm) ready for scanning, the actual scanner speed is 18 pages per minute, and 12 pages can be indexed and released to the system per minute, the net effective throughput is determined by the indexing stage, which is 12 pages per minute. Alternatively, if the document preparation rate is 30 ppm, the indexing and information release yields 20 ppm, but the actual scanner speed is still 18 ppm, the net effective capture throughput is determined by the scanner speed of 18 ppm.

Analyzing Viewing Requirements

Reading document images is very different than reading word processing documents. Unlike word processed documents, in which the clarity and definition of the characters is consistent, document images vary dramatically, both within a single document and among many documents. This places unique requirements on the display that is used to view and read document images.

Each of the individuals involved in the document imaging environment has different needs. Some may closely read and review every document; some may require the ability to annotate the document; some may perform image enhancement upon document image retrieval. Based on document usage patterns gathered in the discovery stage, you should determine the individual and collective needs for viewing features and functionality. The viewing requirements for users who read images a large portion of their day will be more stringent than for users who require occasional access to document images so that they can fax, print or extract one piece of data from the image.

Some key viewing requirements are that the display itself must be large enough to show an entire image. Scrolling between the upper and lower portion of an image is time consuming and cumbersome. And the display must be capable of presenting the image in a readable format. If the image contains handwriting, margin comments or fine print, the display resolution can become a major factor in determining user comfort and productivity. Since document images are primarily black and white, monochrome (black and white) or grayscale monitors provide a clearer image. While color displays provide a vivid viewing environment, there is a tradeoff in image readability. Finally, the display must be fast enough to paint the image on the screen in high resolution as quickly as the operator requests it. Slowly painting the image on the screen leads users to assume that the imaging system is slow.

Not all display subsystems are capable of presenting a readable image to the user, even if that image was properly scanned and enhanced. In addition to the physical capabilities of the display, specialized viewing software is often required to perform image manipulation and enhancement so that the user environment is comfortable and productive. Software features include rotation, scale to gray, zooming, deskewing and annotation.

Analyzing Storage Requirements

The storage subsystem, like each of the other document imaging subsystems, provides a fundamental set of services to the imaging system. And like the others, it must be optimized to work efficiently with image, rather than character, data. Because imaging files are constantly moved in their entirety, imaging is an input/output (I/O) intensive operation. This means that the system can only perform as well as the computer can receive data from and deliver data to the various subsystems. Additionally, large quantities of image files are typically stored for extended periods of time.

The key challenge in analyzing and designing the storage subsystem is to find the proper balance between high-capacity storage and high-performance file delivery. Storage devices that provide high performance tend to have low capacity; while devices with a high storage capacity tend to have low performance speeds. A staged storage environment that provides the right mix and utilization of both types of devices is the key to an optimal storage subsystem environment.

Calculate the Total Annual Storage Volume

First, determine your total long-term storage requirements by calculating the annual and aggregate storage volumes. This calculation must consider the image-retention period, which is the period of time that the images must remain available to system users.

Storage volume estimates should be based on compressed image file size. Compression algorithms result in file sizes that are between 10 and 20 times smaller than the original file. For example, using a compression algorithm that delivers 20:1 compression, a 500KB image file can be compressed to 25KB. Once again, the best way to determine the actual compression for particular document types is to compress sample documents. You can then use that compressed file figure as a reasonable figure on which to estimate annual storage volume requirements for the proposed system.

The formula to determine storage requirements is:

(Number of document image pages) x
{Average compressed image page file size (in KB)} =
Total storage capacity required (in KB)

Bear in mind that storage manufacturers generally claim the unformatted capacity of the media when providing total storage-capacity specifications. A good rule of thumb is to subtract 10 to 15% of the total rated capacity for system overhead and formatting.

Total storage requirements are comprised of on-line, near-on-line and far-line needs. On-line document images are those that need to be immediately available to the end users. High-speed magnetic storage devices are most appropriate to fulfill on-line requirements. Near-on-line storage, which is typically provided by optical jukeboxes, provides users with access to a larger volume of images, but with slightly to moderately slower access times. Document images that are rarely accessed or that are saved for archival purposes can be moved to far-line media, for example, optical disks that have been removed from the storage device.

To determine on-line storage requirements, add the number of images that are coming from the capture subsystem to the number of images that are retrieved and processed during the course of events in a typical day. These statistics are available from the discovery log. You'll need to estimate the number of images that are infrequently requested to determine near-on-line storage requirements. Obviously, the longer an image stays in the system with infrequent retrievals, the sooner it should be migrated to the lowest-cost, highest-capacity storage media available. Far-line storage capacity equals the difference between total storage volume requirements and the combined total of on-line and near-on-line storage requirements.

Determine Appropriate Storage Devices

Imaging systems typically employ a variety of optical and magnetic media, including:

- *Write Once Read Many (WORM):* WORM storage media is a removable optical platter available in a variety of size formats, including 5.25-inch, 12-inch and 14-inch. The most popular format is 5.25-inch, which can store up to 2.6GB of information. WORM disks allow random access, and information can be written in a manner similar to magnetic disk. Unlike magnetic disks, however, once the information is written to the disk, it is permanently etched on the disk and cannot be erased. Unlike magnetic or erasable optical, WORM cannot overwrite the previous version. In this way, WORM can track all revisions to a file written out to disk

- *Erasable Optical:* Erasable or rewritable optical media is also removable, but unlike WORM, enables users to overwrite the previous version of a file. There are many different types of erasable optical media and devices, each of which utilizes a different technology to achieve the same result. Instead of singling out each technology, we have simply grouped them under the category of "erasable optical". Erasable optical capacities range from 100MB to a more typical 1.3GB and higher, depending on the manufacturer and device.

  An optical disk jukebox is a device that can house from 10 to 100 or more optical disks, thus providing greater near-on-line storage capacity. The jukebox makes use of a robotic arm assembly which retrieves the requested disk and places it inside the drive. The jukebox can be configured with a number of drives so that it can be used to both record and play back any of the disks contained in its housing.

- *CD-ROM:* Optical Compact Disk-Read Only Memory (CD-ROM) disks used in computer applications are the same size, shape and dimension as the kind used for audio. Unlike the audio format, however, the data version uses a different set of standards to record and retrieve the information on the disk. The media is capable of storing approximately 650 MB on a disk.

  CD-ROM readers are widely used in the personal computing environment and have achieved the status of "standard" device in current PC configurations. Because of the popularity of both the media and the reader, CD-ROM is highly regarded as a storage media for document images. The low cost of the unrecorded media, approximately $10, further contributes to the strong interest. There are, however, several important issues regarding CD-ROM that must be understood:

  - CD-ROM readers cannot record, they can only play back.

  - CD-R recorders, which can record and play back, are not all standard devices. Some devices "master" disks that can only be read by that device.

  - CD-R recorders are not incremental write storage devices. In other words, they do not write information to the disk as is needed; instead they write information in sessions. In order to write to CD-ROM, all data designated to be written must be stored in a staging area (on another fast hard disk) and be written to the CD in one continuous and uninterrupted session.

  - The number of sessions per disk is limited. Once sessions are used up, so is the disk.

  - Sessions setup in the staging area for recording onto CD-ROM cannot be larger than the capacity of the disk. If the session does not "fit", the disk will be rendered useless, and the operator is typically not notified until the session is concluded.

- *Magnetic Storage:* The most common types of magnetic storage devices (apart from floppy disks) are

IDE, which is the magnetic disk found in most Intel-based PCs, and SCSI hard disk mechanisms. Next to random access memory (RAM), which is the most expensive and also the fastest storage mechanism, hard disks offer the greatest convenience, reliability and speed.

Magnetic storage is the primary storage vehicle for rapid access to images. A single 1GB hard disk can hold approximately 19,000 images. While this may appear to be a large number of files, in some imaging solutions that amount of storage is easily consumed in a single day, or even in an hour.

Certain kinds of magnetic storage systems lend themselves well to both expansion and maintaining data integrity. RAID (Redundant Array of Inexpensive (or Independent) Disk drives is one such storage subsystem. The combination of hardware and software control allows for the potential to create a system that is both fast and provides continuous access despite single or even multiple drive failure. While RAID devices cannot take the place of jukeboxes as they do not support the same capacities nor offer the same cost-per-megabyte for long-term storage, they do provide efficient and high-speed access to large numbers of images that must be on-line for processing.

Storage Management

Imaging systems require large amounts of storage, and by default also require active storage management. The entire storage system must be "staged" in order to provide the maximum tradeoff between capacity and performance. One of the ways to accomplish this is to use high-speed magnetic devices in either a pre-fetching or caching capacity. Pre-fetching images requires the ability to predict what will be utilized during the processing period. These images are then physically moved from a slower high-capacity medium to a high-speed on-line storage subsystem, to which all retrieval requests are directed.

Another means of managing imaging files is to use a high-speed magnetic subsystem as a cache that stores frequently-used images. Unlike a pre-fetch, which involves a batch transfer of files from one source to another, the cache acts a buffer to store only frequently-used images for an extended period of time. RAID devices are commonly utilized in both instances.

Hierarchical Storage Management (HSM)

Effectively managing the storage environment is a task that involves not only the imaging platform, but in some cases the use of higher-level hierarchical storage management (HSM) software. HSM can be used to automatically manage images and data across a storage environment, migrate images to pre-fetch, and move inactive files to the lowest-cost storage device within the storage environment.

Analyze Communications/Network Requirements

As mentioned earlier, imaging is an I/O-intensive process. Because imaging systems are often implemented to assist workgroups or departments in completing work, the I/O intensity affects the communications network of the company. While the temptation to add image traffic to an existing LAN is high, you'll need to evaluate the LAN's ability to shoulder this extra traffic and still perform adequately.

An audit of the installed network to evaluate the condition of the hardware itself, as well as the current performance and utilization of the network, is the first stage of this analysis. A heavily-used network may already be too busy to accommodate additional image traffic, while a slightly-used network may have few visible problems, but might buckle once traffic is increased significantly with document images.

Existing network utilization can be determined with the help of any number of network analysis tools. Some networks are more sensitive than others to dramatic increases in network traffic and actually start to slow down in performance when they reach 30 to 40% of their rated capacity. If your analysis reveals that current network utilization is 15% of capacity, and the allowable maximum is 35%, then 20% is available for image traffic.

Calculate the Impact of Image Traffic

Determining the impact of network traffic is a multi-step, but straightforward, calculation, which begins with the amount of image traffic generated by the capture subsystem upon information release. Network capacity is calculated in bits per second, which means all calculations which result in KB (kilobytes) must be multiplied by 8 to get kilobits.

The formula to determine total annual image network traffic is:

1. *{Number of documents per year} x*
   {Average compressed file size per document (in KB)}=
   Total annual image volume (in KB)

2. *{Total annual image volume (in KB)} x 8 =*
   Total annual network traffic (in kilobits)

This figure can then be divided by the number of work weeks, days, hours and seconds to compute network traffic on a weekly, daily, hourly and per-second basis. By any chance, if this traffic figure exceeds the remaining capacity of the network, the network analysis is complete. To implement document imaging will require that the network be reconfigured, or that imaging be implemented on its own separate network.

Next, you'll need to evaluate peak network requirements by analyzing the network traffic generated by scanning at the peak processing period.

The formula to determine peak image network traffic is:

1. *{Peak processing capture volume (in KB)}/(shift hours) =*
   Image volume per hour (in KB)

2. *Image volume per hour/3600 =*
   Image volume per second (in KB)

3. *{Image volume per second (in KB)} x 8 =*
   Peak capture network traffic (in kilobits per second - kbps)

If peak processing network requirements exceed available network capacity, then corrective measures must be taken.

Third, you'll need to evaluate the network traffic generated by the data extraction stage. Be sure to monitor the steps from scanning to data extraction, so that you identify all the steps involved in indexing. For example, if indexing requires three trips across the network, you'll have to multiply the image volume by three to obtain the total image volume impact on the network.

The formula to determine network traffic for data extraction is:

1. *{Average image file size (in KB)} x*
   {Number of network trips for data extraction} =
   Data extraction image volume (in KB)

2. *{Data extraction image volume (in KB)} x 8 =*
   Network traffic for data extraction (in kilobits)

The fourth step is to compute the network traffic generated by user transactions, as well as the daily traffic generated by image faxes and printed image pages.

The formula to determine network traffic for image transactions is:

1. *{(Number of image transactions per day per operator) + (Number of image faxes per day) + (Number of image prints per day)} x {Average compressed image file size (in KB)} = Image volume per day per operator (in KB)*
2. *{Image volume per day per operator (in KB)}/ (Number of workshift hours) = Hourly image volume per operator (in KB)*
3. *{Hourly image volume per operator (in KB)} x (Number of operators) = Total hourly image volume (in KB)*
4. *{Total hourly image volume (in KB)}/3600 = Image volume per second (in KB)*
5. *Image volume per second (in KB) x 8 = Network traffic for image transactions per second {in kilobits per second (kbps)}*

Finally, you can add the network traffic generated by scanning, data extraction and user transactions to calculate the total image network traffic.

The formula to calculate total image network traffic is:

{Network traffic from peak period capture (in kbps)} +
{Network traffic from data extraction (in kb)} +
{Network traffic for image transactions (in kbps)} =
Total image network traffic (in kbps)

The result of this last equation should then be compared to the total capacity available on the network to determine whether additional network capacity is required. If additional capacity is required, you can consider implementing a separate LAN for the capture subsystem, enhancing the current network capability, or develop a series of work arounds that divert network traffic out of the busy lanes required by the imaging system.

Analyze Printing Requirements

The perception that imaging does away with paper, and therefore does away with printing is largely mistaken. Rather than eliminate paper entirely, imaging manages paper documents in a better way. The net result of that enhanced management is that the document becomes more accessible, and consequently more easily and more frequently printed.

However, printing images is not the same as printing a word-processed letter. As mentioned earlier, images are large objects that are reduced in size via software compression. In order to print an image it must be decompressed. A decompressed 200dpi image is approximately 500KB in size. Thus a 10 page report is 5MB in size, which will take a considerable amount of time to print. If the need arises to print several reports per hour, the printer may simply not be fast enough to perform the task.

The second critical issue for printing images is the point of decompression. If the image is decompressed at the local viewing station and is being printed to the locally-attached printer, then that workstation will remain busy and inaccessible to the user until the printing process is complete. Conversely, the image can be locally decompressed and then sent to a network printer. However, if there is a substantial volume of print activity, this can easily inhibit network performance.

There are two basic architectural alternatives:

1. Send compressed images to a print server, which then prints to an attached printer;
2. Use a print accelerator card in the printer and send compressed images to it;

To achieve maximum throughput, both alternatives can be implemented.

Specification and Design: Configuration Criteria

After analyzing the impact of document images on each of the computing subsystems, the next task is to specify product and design requirements. You'll need to determine two categories of specifications:

1. Product specifications to identify the types of products that meet system parameters;
2. Design specifications to outline the manner in which the system must be configured in order to meet performance specifications.

While to some extent, the performance of each subsystem may be device-specific, the functionality of the subsystem depends heavily on overall configuration characteristics. That means that you'll need to continually balance product and design specifications. For example, the actual throughput of the scanning device may be lower than anticipated. A faster scanner may deliver greater throughput, but the faster scanner may require more index stations and operators to keep up with the scanner output.

While product specifications and design considerations always go hand in hand, the first priority of the specification stage is to extract minimum product specifications. So, for example, if the capture subsystem must release 5,000 images per work shift, then a minimum capture performance of 5,000 is required, regardless of other design issues.

As you review the subsystem specification parameters a combination of performance and feature requirements will emerge. For example, the daily document image volume indicates the level of performance that you'll need from the capture subsystem, i.e. the actual scanner throughput rate, with peak processing volumes setting the upper portion of that performance spectrum. The analysis also identifies specific features that you'll need, for example, image enhancement built into the scanner. Finally, the analysis indicates potential "throughput inhibitors", such as the time spent on document preparation, the number of document types, or the amount of time required to complete the index fields on a document. These inhibitors can be addressed via a combination of product and design specifications.

The statistics gathered from each of the previous phases should drive the specification process. For example, the analysis of storage requirements provided two critical capacity calculations:

1. The total long-term storage volume requirements, which are based on the annual document volume and the retention period required for that volume. This figure then drives the specification of the storage device. For example, it can be used to determine the number of slots required in a jukebox, which in many cases will serve as the primary storage repository.
2. The on-line storage requirements for the work processing period. The on-line capacity requirements, combined with the number of transactions performed and the time allowed for the transaction to be completed, determine the size and speed requirements for the on-line storage environment. The use of cache or pre-fetching is dictated by the nature and amount of transactions that need to be performed within a specified time period.

Design Specifications

Imaging system design starts with the document and ends with the process. In other words, the document itself determines a great deal of the imaging system functionality while the process associated with the document drives the operational design of the system. The process also drives the capabilities required from the imaging platform.

Each subsystem has its own purpose. The capture subsystem must deliver indexed, high-quality images within the allotted processing time frame. The storage subsystem's function is to store images in the most economical manner over the long term, and in the highest-performance manner during the processing cycle. The communications subsystem is the highway, over which images are delivered to their destination, and so on. Each subsystem must be fully optimized in order to provide the necessary services and performance. The first design objective is to make sure that each subsystem can perform as expected within the subsystem's operational domain. The next step is to ensure that all these subsystems work well together. That's where the imaging platform comes in.

The job of the imaging platform is to provide the ability to integrate those subsystems within the standard computing environment, and to effectively manage document images across the entire system. In iterative fashion then, the facilities which provide the core level or imaging infrastructure first are fine tuned, integrated and debugged. Once these facilities are in place, at the very least on paper, the design process can progress to the next level, which is the creation of the management application. Image management applications are created utilizing development tools like Microsoft's Visual Basic, in conjunction with tools provided with the imaging platform.

Matching Application Requirements with Technology Solutions

Not only did the discovery process yield information relevant to the design of each subsystem, but it also provided a wealth of information related to the requirements for the management application. By observing the functional flow of documents through the department, you have discovered the operational issues related to the documents and how they are processed. Among the observations: the type of work performed at each location; the movement of work among different locations; the standard databases, spreadsheets and other computing tools used in the process; and the interaction of the document process with these and other applications. Each of these details is critical in determining the application's work management and transaction processing requirements.

One of the primary considerations regarding the management application is to ascertain the need for, and level of, workflow automation. The group-processing nature of many document imaging applications leads many to assume that workflow is always a required part of the imaging solution. But there are actually various ways to manually stage a work process and configure an imaging system to provide some level of work routing and management without using explicit workflow tools or applications.

For example, incomplete folders can be routed to a separate location where they are accessed by users dedicated to processing these folders. After they are completed, they can be moved to join other folders. Alternatively, a hierarchical storage manager can automatically migrate work batches that must be completed within a certain time frame to a server located in the department where users can finish the work before the deadline.

What cannot be easily automated without workflow software are complex work processes that require sophisticated work management and distribution features based on predefined conditions.

These highly-structured work processes can benefit immensely from the deployment of workflow. Not only can workflow facilitate the management and tracking of work among users, it also can automate the imaging system itself. An example is the use of workflow software to automate a complex capture and indexing process so that a steady stream of already-indexed output flows from the capture subsystem.

Another critical design issue is the level of required integration between the imaging application and other applications. Users may require access to other desktop applications, such as word processing or spreadsheet applications, along with the imaging software, or the imaging system may need to integrate with an existing host-based data processing application. Imaging systems are commonly integrated with database applications, and it is generally safe to assume that some level of integration is required.

While integration is nearly almost possible, the issue is the degree and ease of the integration. That depends, to a great extent, on the imaging platform selected. Virtually all platforms support screen-level integration, which is the ability to run a host session on one part of the operator's viewing screen while the imaging application runs on the other. More tightly-coupled integration can be achieved by creating a program, using tools like Visual Basic or PowerBuilder, that links the imaging and the legacy application together with one common interface. Deeper levels of integration are also possible but depend largely on the imaging platform's support for APIs (Applications Program Interfaces) and other integration interfaces.

A number of imaging platforms, through their support of Microsoft's ActiveX technology, can be integrated with other desktop applications like e-mail and word processing. Other applications compliant with the ODMA (Open Document Management API) standard can also be supported as well as applications compliant with Microsoft ODBC and SQL.

The last major factor regarding application design is the level of customization needed, at both the application and system level. Several years ago, it was quite common to have system-level customization because certain performance goals could not be met with standard devices or the standard devices needed were not supported by the platform in use. This type of customization is far less common now.

Today, customization is primarily conducted and required at the application level. While one company's accounts payable application works in a similar way to another accounts payable application at the system level, the nature of the documents processed may vary considerably, and therefore will require some application customization. The nature and degree of required integration with existing applications also varies from company to company, so those links will have to be customized as well. Again, the degree of support for customizing applications, as well as the customization techniques employed, vary considerably among imaging platforms.

Implementing Imaging within the Technology Environment

The Software and Networking Environment

Imaging system software platforms can operate in nearly all types of standard computing environments. Imaging systems are most commonly built around Windows 3.x clients, with either Windows NT, OS/2, or UNIX servers. A number of imaging platforms now support Windows 95 and Windows NT clients as well. All common networking environments, including OS/2, Windows NT, Novell NetWare and Banyan Vines, are supported by most imaging applications.

Most imaging platforms rely on an image viewer and an image server to provide imaging services. Some imaging systems store both images and indexes on the same server; others store images on one server and use third party databases to store the index information. The database then maintains a permanent link to the images in the image server. Some third party databases can store the image and the index in the same database record. These types of databases are called BLOB (Binary Large OBject) databases.

The viewer application enables users to call up and display document images. Some viewers include additional features, such as local scanning, fax and print support, and document annotation. A few imaging platforms provide their own user interface with viewing features implemented as part of the desktop. In general, image viewers provide limited functionality; it is the server that does the heavy lifting. The server oversees the administration and management of the entire system, including the management of the imaging database, user access and security, and access to subsystem services.

The Hardware Environment

The computing hardware environment must be fairly robust to support document imaging applications. The fact that image compression and decompression is primarily done in software dictates a strong processor at the client workstation, typically an Intel 486-based PC. Random Access Memory (RAM) is also at a premium in imaging environments, which generally require a minimum of 16MB.

Server hardware must also be powerful, both in terms of processor speed and memory. The server should also possess strong expansion capabilities. Many of the external devices utilized in document imaging are built around the SCSI (Small Computer System Interface) standard which means that the server must be configured to support SCSI, and possibly numerous SCSI expansion slots.

Implementation Guidelines for Successful System Deployment

As we mentioned at the outset of this guide, successful imaging systems start with the business benefit. Business benefits are much easier to identify and achieve in discrete application areas suffering from habitual document control problems than in the enterprise as a whole. While many imaging systems are capable of scaling up to support more users in an organization, a single imaging application rarely scales up to cover the entire enterprise.

The more likely scenario involves the gradual expansion of individual applications, which occasionally need to share information with each other. It is unlikely, for example, that the customer service department needs regular access to all of the images in the accounts payable department, and vice versa. Determining the throughput and processing boundaries of an imaging environment in one discrete application is infinitely simpler than performing that analysis for an entire corporation. In other words, the ability to clearly define a single application is more important, and more likely to lead to successful system implementation, than defining all possible applications at the same time.

The successful deployment of an imaging system requires a mix of resources, many of which may not be available internally. Document imaging is a field with its own specialized vocabulary and a rapidly changing technical environment. For this reason, a mix of internal and external imaging specialists is the most desirable course of action. The external resources have the familiarity with imaging specific applications; the internal resources have the familiarity with the existing computing infrastructure. If selected and managed effectively, the combination should provide a winning formula.

The Next Step

The data collection and analysis discussed in this guide should precede any evaluation of a specific imaging solution. With this analysis complete, the next step is to evaluate vendors, resellers and integrators to determine who best understands your business problem, and which solutions best meet your business and technical requirements. The work you have performed during this analysis will tell you a great deal about the issues involved in selecting the right solution for your needs, and will help guide the eventual evaluation, selection and implementation stages.

CASE STUDY 1—Airport Tames its "Paper Traffic Control"

When you think of airport traffic jams, you probably picture hundreds of cars lined up outside the terminals, or dozens of airplanes stacked up waiting for takeoff. But the most daunting traffic jam faced by a typical airport is the endless stream of paper that floods the facility every day. In one case, San Francisco International Airport (SFO) used high-end computer technology to help tame its "paper traffic control" problem.

As with most large organizations, SFO made thousands of copies every week. Important documents were kept in large filing cabinets, the sheer number of which quickly overwhelmed several sizable storage rooms. Over time, documents were lost or misfiled. They were also vulnerable to flood, fire and other elements. With SFO set to embark on a $2.5 billion capital improvement program, it was evident that the old filing system would have to be replaced.

"We were afraid of what might happen when the paperwork for a $2 billion project was added to a system that was already strained," said Marilyn Williams, program services manager for SFO. "The number of documents we deal with daily is incredible. The filing system had to be upgraded with something that would cut our need for storage space and yet be simple to use."

SFO's Master Plan for solving its worsening paper problem called for a new filing system that had to be paperless, eliminate the need for numerous filing cabinets and be fast enough to handle everyday volume. With today's document imaging systems, thousands of documents can be stored on one tiny disk, freeing-up much of the space needed for a filing system and saving money in the long run.

SFO was required to solicit bids for the Master Plan contract. Because of the overall simplicity of its system and the proven reliability of the proposed hardware and software, SFO selected SpeedScan of Skokie, IL. "We knew that filling the requirements of the Master Plan would be a challenge, but we knew we could design a filing system that met all their demands," said Steve Gilford, SpeedScan's director of sales. The company's proposal specified a Panasonic high-speed production scanner (KV-SS50), laser printer (KX-P6100) and optical storage drive (LF7300A). The peripherals are centered by a Compaq Pentium PC.

"I think SpeedScan's plan was the best possible solution for us," said Williams. "Their system met all our basic requirements while using components that could be adapted to our existing network. There are a lot of scan-to-file systems available, but many use proprietary hardware. They get the job done, but don't offer the overall flexibility we needed."

The new system scans incoming documents and stores them on the Panasonic optical drive. With an amazing 1.5 gigabyte (GB) of storage space, the drive can store up to 30,000 documents safely and securely on one 5 inch optical disk with a guaranteed life of 10-15 years. With the use of WORM (Write Once

Read Many) disks, delete functions are disabled to prevent files from being accidentally erased, dramatically reducing SFO's need for space-consuming file cabinets.

With incoming documents scanned into the system, SFO immediately realized a cost savings by reducing the number of copies it makes. And with SFO's e-mail system, it takes only seconds to simultaneously distribute important documents to employees' desktops.

With SFO's new file storage system, important communications between city planners and contractors were scanned and stored by date. "We felt, from a legal standpoint, it was important to get our filing system under control before we tackled the massive reconstruction of SFO," said Williams. The potential for future liability was lowered; since files are never physically removed from the system, loss or misfiling of important documents is virtually eliminated. If any is filed at a future date, SFO is confident that the information it needs to protect itself is securely filed away.

"We knew that the correspondence between city planners and contractors could be important in the future," said Williams. "The possibility of legal action in a project of this size is so great, we wanted a filing system that was not only able to store important documents, but was also secure enough to prevent loss of documents."

The Master Plan is approximately 65% complete, and all incoming correspondence is now scanned into the system. "We have to tackle the backlog of information that's sitting in our filing cabinets," said Williams. "It'll take us a couple months to get to it all, but when that's done, we'll save so much room. I'd have to say that we're very happy with the system."

Systems like SFO's Master Plan are becoming more common every day. Just as large international airports and Fortune 500 companies are finding their old filing systems inadequate, so, too, are smaller organizations recognizing the shortcomings of their costly paper filing cabinets.

CASE Study 2—Document Imaging Helps Children at Risk

Document imaging technology is not only helping businesses run more efficiently, it's also enhancing government's ability to defend its most vulnerable citizens. Protecting America's children from deadbeat parents has become one of the country's most vexing social issues. Because of overcrowded courts, many of these children go without basic necessities like adequate food, clothing and shelter. In an attempt to improve enforcement of court-ordered child support and visitation, the family court of Sumter County, SC has added a state-of-the-art document imaging system from Panasonic Computer Peripheral Company.

Sumter County's paper file system was unable to handle the growing demands of a modern family court system. The sheer number of documents to be filed and stored overwhelmed the system. The Family Court of Sumter County processes 107,000 child support payments annually, adding about 60,000 new documents each year to the ever-increasing volume. In addition, paper files can only be handled by one person at a time. And since court workers cannot work on files simultaneously, this antiquated process slows the mammoth system to a crawl.

Unlike a civil or criminal case that's closed at its conclusion, a family court case can remain open for up to 18 years, creating even greater problems for the filing system. With the overwhelming backlog and limited access to information, it has been reported that deadbeat parents find it easy to fall between the cracks.

"Our first responsibility is to protect the children of the county," says Larry Fudger, project manager for Sumter County. "The problem is that even though the law demands immediate action when a parent is late, it's difficult for the court to get timely information. It can take us weeks or even months before we can take action in a case."

Like many court systems across the country, Sumter County needed a way to get information to caseworkers quickly. County officials decided to adopt an electronic filing system that scans 40 pages per minute and stores documents on 5 inch disks with an incredible 30,000-page (1.5 gigabyte) capacity, retrievable at the touch of a button. The system's fast scanning maximizes filing efficiency, saving time and allowing caseworkers to focus on their main task of aiding children. It also features quick retrieval, reducing the time it takes to locate files or search for off-site, misfiled or lost documents. As a result, the processing time for one case has been reduced from weeks to a single day. From their desks, court workers can now gain access to any file in the court, and can quickly determine if parents are in compliance with child-support and visitation orders. The new streamlined system lets caseworkers get involved as soon as a problem occurs.

The Sumter County program was funded by the U.S. Department of Health and Human Services with money that was earmarked for use in enforcing child support and visitation rulings.

From America's largest corporations to its smallest municipalities, the need for quicker, leaner, more efficient filing systems is being met. As prices come down and more organizations realize the benefits of such systems, the demand for computer-based file storage is growing rapidly.

The time wasted by digging through mounds of paper to locate an important document can no longer be afforded by many organizations. With new file storage technology, the file room can now be at your fingertips, organizing data more productively than any paper file system. What once took hours to do using precious office space and overworked resources can now be accomplished efficiently, reliably, and conveniently.

Glossary

American National Standards Institute (ANSI)—A standard-setting, non-governmental organization, which develops and publishes standards for "voluntary" use in the United States.

American Standard Code for Information Interchange (ASCII)—The most popular coding method used by small computers for converting letters, numbers, punctuation and control codes into digital form. Once defined, ASCII characters can be recognized and understood by other computers and by communications devices. ASCII represents characters, numbers, punctuation marks or signals in seven on-off bits. A capital "C", for example, is 1000011 while a "3" is 0110011

Application—A broad and generic term for any software program that carries out a useful task. Word processors and graphics programs are applications.

Application Program Interface (API)—Generic term for any language and format used by one program to help it communicate with another program. Specifically, an imaging vendor can provide an API that enables programmers to repackage or recombine parts of the vendor's imaging system, or integrate the imaging systems with other applications, or to customize the user interface to the imaging system.

Architecture—Refers to the way a system is designed and how the components are connected with each other. There are computer architectures, network architectures and software architectures.

Archive—A copy of data on disks, CD-ROM, mag tape, etc., for long-term storage and later possible access. Archived files are often compressed to save storage space.

Association for Information and Image Management—Trade association and professional society for the micrographics, optical disc and electronic image management markets.

Automated retrieval—Using a computer to identify and locate a stored image of some kind. Generally requires the use of key words or codes in an indexing scheme.

Backbone—The part of the communications network which carries the heaviest traffic. The backbone is also that part of a network which joins LANs together—either inside of a building or across a city or the country.

Backfile conversion—The process of scanning in, indexing and storing a large backlog of documents on an imaging system.

Background—(1) The simultaneous, non-interrupting, execution of an automatic program while the computer is being used for something else. (2) The portion of microfilm that doesn't have anything recorded on it.

Backup—A duplicate copy of data placed in a separate, safe "place"—electronic storage, on a tape, on a disk in a vault—to guard against total loss in the event the original data somehow becomes inaccessible.

Bar Code—A system of portraying data in a series of machine-readable lines of varying widths. The "UPC" on consumer items is a bar code. In document management, a bar code is used to encode indexing information.

Batch—A group of one or more documents of the same document class (such as invoices or tax forms).

Batch processing—Conducting a group of computer tasks at one time, instead of steadily throughout the day.

Bilevel—A binary scan that assigns each pixel an attribute of either black or white—no gray tones, no colors.

Binary digit (bit)—Represents the binary code (0 or 1) with which the computer works. A number of bits together are used to represent a character in the computer.

Binary Large Objects (BLOB)—The ability to embed large binary objects (images) as part of a character database record.

Bitmap—Representation of characters or graphics by individual pixels, or points of light, dark, or color, arranged in row (horizontal) and column (vertical) order. Each pixel is represented by either one bit (simple black and white) or up to 32 bits (fancy high definition color).

Bitmapped image—Representation of image data where each pixel has a corresponding memory element.

Black line—A positive image, black on a clear or white background. Opposite of "white line." Also known as a negative image.

Cache (Memory & Magnetic)—Small portion of high-speed memory used for temporary storage of frequently used data. Reduces the time it would take to access the data, since it no longer has to be retrieved from the disk.

CCD—Charge coupled device. A semiconductor memory device within which stored information circulates rather than remains in fixed position. Also used to describe the set of circuits which commonly make up the "eyes" of a scanning device, or video camera.

Character recognition—The ability of a machine to read human-readable text.

Compact Disc Read Only Memory (CD-ROM)—A data storage system using CDs as the medium. CD-ROMs hold more than 600 megabytes of data.

Compression—A software or hardware process that "shrinks" images so they occupy less storage space, and van be transmitted faster and easier. Generally accomplished by removing the bits that define blank spaces and other redundant data, and replacing them with a smaller algorithm that represents the removed bits.

Computer Output to Laser Disk (COLD)—Technique used to transfer computer-generated output to optical disk.

Consultative Committee for International Telegraph and Telephone (CCITT)—International organization that develops international communication standards.

Data compression—Reducing the amount of electronic "space" data takes up. Methods include replacing blank spaces with a character count, or replacing redundant data with shorter stand-in "codes." No matter how data is compressed, it must be decompressed before it can be used.

Data decompression—The regeneration of a bit-map from a compressed representation.

Data transfer—The movement of data inside a computer system.

Database—Data that has been organized and structured in a disciplined fashion, so that access to information of interest is as quick as possible. Database management programs form the foundation for most document storage indexing systems.

Decompress—To reverse the procedure conducted by compression software, and thereby return compressed data to its original size and condition.

Departmental imaging system—A multi-workstation imaging system typically used by at least three and usually by more than 25 people in a workgroup or department.

Deskew—To straighten out a crooked image. Improves OCR accuracy and reduces image file size.

Despeckle—An image processing (clean-up) operation that removes random specks, called flyspecks or pepper, from an image to improve legibility, OCR accuracy and to improve compression.

Device driver interface—A function in an API that allows developers to write drivers that write directly to hardware and do not rely on interrupts.

Device drivers—Small programs that tell the computer how to communicate with a particular types of peripheral devices.

Diffusion—A type of halftone that simulates grayscale or color. The dots in the image are distributed in a random pattern, resulting in increased image detail and better texture.

Digital Scanner—Optical reader that scans and converts images into digital form.

Disk array/Disc array—Combining redundant disk or disc drives for more capacity, or for disaster recovery.

Disk array controller—Acts as a manager between the host and the drives. Comprised of a main computer module, channels for each drive, and a host channel for each host input. Adding memory modules increases performance.

Disk cache—Place in memory where frequently recalled data is stored while you're working. It makes retrievals much faster.

Disk mirroring—A fault-tolerant technique that writes data simultaneously to two hard disks using the same hard disk controller. The disks operate in tandem, constantly storing and updating the same files. Mirroring alone does not ensure data protection. If both hard disks fail at the same time, you will lose data.

Document retrieval—The ability to search for, select and display a document or its facsimile from storage.

Document staging—In a document retrieval from an optical jukebox, the process where the image is fetched from the server by the software, and stored on the user's local PC until it is used.

Dots per inch (dpi)—A measurement of output device resolution and quality. Measures the number of dots a printer can print per inch both horizontally and vertically. A 600 dpi printer can print 360,000 (600 by 600) dots on one square inch of paper. More dpi means higher resolution and greater detail.

Drum scanner—An ultra-high resolution scanner, often used for scanning transparencies. The slide is attached to a rapidly spinning drum, which turns under the scan head. It moves very slowly across the image, resulting in very close lines of dots, thus high resolutions.

Electronic Data Interchange (EDI)—An electronic communications standard which connects business trading partners for conducting contract negotiations, sales, invoicing and collections.

Flat-bed Scanner—Device for scanning that has a flat surface for input material. Generally used for scanning bound material.

Gateway—Conceptual or logical network station that serves to interconnect two otherwise incompatible networks, network nodes, subnet-works or devices. Gateways perform a protocol-conversion operation across a wide spectrum of communications function or layers.

Handprint Character Recognition (HCR)—The ability of a computer to read handprinted characters, not generated by machine.

Hierarchical File System (HFS)—In DOS, the file management system that allows directories to have subdirectories, and sub-subdirectories. In Macintoshes files may be placed into folders, and folders to be placed within other folders.

Indexing—A method by which a series of attributes are used to uniquely define an imaged document so that it may later be identified and retrieved.

Intelligent Character Recognition (ICR)—Advanced form of OCR technology that may include capabilities such as learning fonts during processing or using context to strengthen probabilities of correct recognition.

Jukebox—A device that holds multiple optical disks and one or more disk drives, and can swap disks in and out of the drive as needed. Same as an autochanger.

MAPI (Messaging API)—A Microsoft published API that separated the client from the server functionality, allowing various clients, like mail front ends, word processors, spreadsheets, etc., to access the messaging capabilities of back-end mail servers, such as Microsoft Exchange Server.

OCR indexing—Technique that recognizes lettering, numerals and other characters and converts them to ASCII (attributes of a document) for subsequent retrieval.

On-line—Data that is available on a primary storage device so that it is readily accessible to the user.

Optical Character Recognition Reader (OCR)—The ability of a scanner with the proper software to capture, recognize and translate printed alphanumeric characters into machine readable text. Most OCRs work by using either Pattern Matching or Feature Extraction.

Optical disk—A storage device that is written and read by laser light. Certain optical disks are considered Write Once Read Many (WORM), because data is permanently engraved in the disk's surface either by gouging pits (ablation); or by causing the non-image area to bubble, reflecting light away from the reading head. Erasable optical drives use technologies such as the magneto-optic technique, which electrically alters the bias of grains of material after they have been heated by a laser. Compact disks (CDs) and laser (or video) disks are optical disks. (When referring to CD technology, the spelling disc is used. In all other cases, the spelling disk is used.)

Redundant Arrays of Inexpensive or Independent Discs (RAID)—A storage device that uses several optical discs working in tandem to increase bandwidth output and to provide redundant backup.

Rewritable optical disk—Optical disk on which data is recorded. The data in specified areas can subsequently be deleted and other data can be recorded.

Tagged Image File Format (TIFF)—A bit map file format for describing and storing color and gray scale images.

Visual Basic Extensions (VBXs)—Add on Programs (written in C or C++) to the Microsoft's Visual Basic development environment that perform specific functions, such as putting a menu on a screen, or deskewing images). VBXs can be combined to build customized document imaging application modules.

Write Once Read Many (WORM)—Optical storage device on which data is permanently recorded. Data can be erased, but not altered, and no additional data can be added.

RHEINNER GUIDE

This Technology Guide is one of a series of guides, written by The Rheinner Group and published by ATG, designed to put complex document management, imaging, and workflow concepts into practical and understandable terms. Each guide provides objective, non-biased information to assist in the internal education, evaluation and decision making process. This Technology Guide, as well as the other Document Management, Imaging, and Workflow Technology Guides in the series, are available on ATG's Web Site.

www.techguide.com

A joint venture

ATG
The Applied Technologies Group

Tel: (508) 651-1155
Fax: (508) 651-1171
E-mail: info@techguide.com

The Rheinner Group

Tel: (617) 741-8100
Fax: (617) 741-5885
E-mail: info@rheinner.com