Sizing a Document Management System:
Image Size Estimates for All Types of Digitized Documents

INTRODUCTION

This paper provides managers with document size estimates for use in discussing, planning, designing, and implementing a document management system. The average computer file sizes for many types of scanned (digitized) documents are listed and explained.

This paper refers to the tables in white paper 22009, Computer Storage Requirements for Various digitized Document Types.

DESIGN OF ESTIMATES

Round Numbers

All of the estimates are chosen to be representative of the file sizes of scanned document images and to produce easily useable, round figures, when multiplied. For example, 50 thousand bytes (50 Kilobytes) is close to the average size of scanned pages and also yields an estimate of exactly one million bytes (a MegaByte) for 20 pages, exactly one billion bytes (a GigaByte) for 20 thousand pages, and exactly one trillion bytes (a TeraByte) for 20 million pages.

Over-Estimation

These estimates have been chosen to tend toward over-estimation rather than under-estimation of storage requirements. The estimates therefore provide a small safety margin. It is best to make certain that all estimates and estimating procedures tend toward conservative estimates. Then, when all assumptions are factored in together, aggregated, the overall estimate is conservative (safe).

Metric Round Numbers

The United States became a metric country in 1866 (Act of July 28, 1866; 14 Stat. 339) when the US customary measures were defined in metric terms. For example, an inch is defined to be exactly 25.4 millimeters. While an inch, as a measure, is thus hard metric, it is not a round metric number. When physical items are hard metric, the quantity and size of the items are in round metric numbers.

In this paper, US customary measures are converted to metric sizes and quantities by applying the above methods for constructing round numbers and ensuring a slight over estimation. These methods are applied to the quantities and sizes of documents listed once for the US customary units, and a second, independent time for metric units. The resulting metric measures are then chosen to be appropriate to the corresponding US customary measures. ('~' is used to indicate 'close', '~~' is used to indicate 'not so close, but similar'.) This means that the conversion process produces similar conceptual values, but almost never produces numerically or physically equal values (e.g. – 40 degrees Celsius = – 40 degrees Fahrenheit) for the quantity or size of a given item.

Many scanning applications in metric settings use hard US customary measures for image resolution, e.g. 200 dpi (dots per inch) (about ~ 8 dpmm – dots per millimeter), 300 dpi (~12 dpmm), 400 dpi (~16 dpmm), 600 dpi (~24 dpi). The most common metric scanning (or printing) resolution is 2540 dpi (exactly 100 dpmm).

Standard Estimates

If everyone uses the same estimates, it is easier to discuss and compare document imaging systems. Managers can also benefit from reports and articles describing previously implemented systems.

Because the estimates are industry standard, less time can be spent evaluating estimating methods, and more time can be spent understanding how the system will be used and whether the system design will accommodate the planned use.

De Facto and De Jure Estimates

De facto means 'by common usage', in the same way a dictionary records the way in which a language is actually used, rather than being the rule by which the language is written and spoken. These estimates are defacto estimates, estimates commonly used in the document management industry.

De jure means 'by some agreement or authority'. An example is the ISO (International Standards Organization) standards process where there are policies and procedures for accrediting standards setting bodies and for the functioning of those bodies. Metric paper sizes are an example of an ISO standard.

Average?, What About My Documents?

Making the assumption that your documents are similar to the industry average documents usually produces very small variances. Because the cost of storage is very low as a percentage of overall system cost, and is dropping rapidly, an error of a few percent in an estimate has very little effect on the overall system cost. If round estimates speed up the understanding and discussion process, the benefit of rounding far out-weighs the cost of the slight variances.

After one percent of the documents have been scanned into a system, an actual average page image size can be calculated. This actual average page image size will almost always provide the small correction necessary to adjust previous estimates. This is the system sizing method used in almost all system implementations.

Estimating

Managers need ball park figures to size a document management system, to identify what system elements are bigger (or smaller) than a breadbox (system or system component). Managers need a rough order of magnitude (ROM) estimate. An order of magnitude is a power of 10 so that a rough order of magnitude estimate (ROM) is one in which the largest reasonable estimate is about ten times the size of the smallest reasonable estimate.

The accompanying list of document sizes is intended to assist in creating ROM estimates of storage requirements. Possible system decision outcomes of making these estimates are: "There will be no problem (the system is a little too large).", "We will never be able to afford the system.", and "We budgeted the right amount.".

Variance Management

Managers manage variances, the difference between carefully made estimates (or budgets) and actual figures. These image size estimates are intended to initiate the process of variance management.

Compression

All of the figures given for scanned images are for compressed file sizes, unless otherwise noted. All imaging systems

compress their image files for storage and transmission. Compressing removes the redundancy from the files, making the files smaller. These compressed page files have an average size of approximately 50 thousand bytes per page. Rarely is a given compressed page file exactly 50 thousand bytes.

What About Different Resolutions?

Image files created by scanning at 200, 300, and 400 dpi (dots per inch) all have the same information content as the original image. The higher resolutions merely increase the redundancy in the image file. It is this redundancy that compression removes. In general, higher resolution scans of an image are slightly larger than lower resolution scan of the same image because higher resolution scans pick up more noise (pseudo-information). (Noise is something like digital dirt on the image.)

This variation between the compressed image sizes of different resolutions is within the variation range of document image sizes in general. In almost all cases, measuring the actual sizes of the first one percent of scanned images will easily adjust for this variation without requiring significant system changes.

DESCRIPTION OF THE TABLE OF STORAGE REQUIREMENTS THAT FOLLOWS

The following is a narrative of the information in the table of document type storage requirements that is included below, as a 4 page pullout section.

SCANNED LETTER SIZE PAGES

Legal Size

Legal size page images are not much larger than letter size page images when scanned and compressed. The pathological case is the eight and one-half by twenty-five inch contract set in six point type, created to be true to the statement that the entire contract is on just one page (almost always two sided). Even in these cases, a good system design can be arrived at using letter size page image estimates. Measuring the size of actual scanned pages after the first one percent of the legal size documents have been scanned will produce the same high level of accuracy produced by measuring the size of the first one percent of any type of scanned document images. With this foundation of industry de facto standards, adjustments can be made for even the worst pathological cases, and any desired degree of estimating accuracy can be achieved.

Standard Record Storage Cartons and Fan-Folded Computer Output

A standard records storage carton (box) is about 12 inches wide by 15 inches long by 9 1/2 inches deep (300 mm x 375 mm x 235 mm). It is designed to store letter size documents in manila folders against the 12 inch (300 mm) side and legal size (8 1/2 by 13 or 14 inches) (210 by 325 or 350 mm) documents in legal folders against the 15 inch (375 mm) side. The standard fan-folded, greenbar, tractor fed, 11 by 14 inch (275 mm x 350 mm), computer paper can be placed flat on the bottom of the standard records storage carton. In all three cases, the total computer storage required to store the scanned and compressed images of the documents in the box is the storage required to store 2,500 letter size page images.

By counting a standard record container that contains about 2,000 legal pages, as having 2,500 sheets of letter size documents, the effect of the slight difference between legal and letter pages can be further reduced.

Because the mainframe style programs that produce fan-folded greenbar output use very simply typography, the pages are simple and compress well, to about the same size as an average letter size page. Fan-folded documents are on good paper and must be handled carefully or they quickly become unmanageable. For these reasons, the 9 or so inches (225 mm) of fan folded documents that will fit in a standard records storage carton are fairly dense and constitute about 2,500 sheets.

The Difference Between Pages and Documents

Unfortunately, the document imaging industry has blurred, and then finally eliminated, the difference in meaning between the words 'page' and 'document'. This happened because there is a desire to make every system seem as large as possible, so every stored page is said to be a document.

To recover the use of the word 'document', always separately list the number of documents, the number of pages, and the average number of pages per document. This will allow a discussion of pages and documents to continue without losing track of whether or not pages are the same as documents.

Simplex and Duplex

Simplex means one-sided pages and duplex means two-sided pages. Simplex is always assumed. In the same way that pages and documents can be confused, so can the count of two page images on one duplex page. To avoid this, always separately list the number of pages and the number of page images.

Bits and Bytes

Computer storage is given in bytes and transmission speeds are given in bits. To avoid confusion, always spell out the word bit and byte in all planning documents. It is very common for bits to turn into bytes and for bytes to turn into bits during discussions and phone conversations. This results in plans that are either eight times too large or eight times too small (or too slow, or too expensive).

Pages, File Cabinets, Boxes, and Linear Feet

Using an estimate of 2,500 pages per file drawer and four file drawers per file cabinet, one can estimate that scanning one four drawer file cabinet full of documents (ten thousand single sided pages) will fill one CD ROM disc. Similarly, the scanned contents of two file cabinets will fill one GigaByte of magnetic disk storage.

With these figures, a simple count of the file cabinets in an organization will produce an estimate of the amount of storage required. At an even coarser level, if two file cabinets are assumed for each employee, the number of GigaBytes required is equal to the number of employees.

Files in file cabinets and on linear feet of open shelving (4 linear feet ~~ 1 meter) are assumed to have some open space for file growth and for ease of access. The number of pages estimated for these two storage methods takes this into account.

Standard record storage cartons (boxes) are assumed to be more tightly packed because documents in boxes are placed there for storage. Access to documents in boxes is assumed to be less frequent than to documents stored in active file cabinets and on open shelves. The estimates also take these assumptions into account.

An Adjustment Factor

If your storage facilities are tightly packed or are otherwise different than these estimates, you can apply a correction factor to adjust for you facility's differences. For example, a

tightly packed facility might have an adjustment factor of 1.1 because there might be ten percent more pages per linear foot or drawer than in the industry standard density.

The adjustment factor is not required for ROM (ballpark) estimates. Also, the adjustment factor is easy to apply at any stage in the process. If after weeks of work, a storage estimate of 100 GigaBytes is arrived at, the decision to use an adjustment factor can be made, and the storage estimate can simply be adjusted to 1.1 times 100 GigaBytes yielding an estimate of 110 GigaBytes.

Box and Microform Sizes

If you have double length boxes, such as 12 x 30 x 9 1/2 inches (300 mm x 750 mm x 235 mm) instead of the more standard 12 x 15 x 9 1/2 inch boxes (300 mm x 375 mm x 235 mm), simply multiply the number of pages per box by 2.

If you have 200 foot (60 m) rolls of microfilm instead of the listed 100 foot (30 m) rolls, simply multiply the number of pages by 2.

A 100 ft. (30 m) role of 16 mm microfilm of 24X images actually has closer to 2,400 images rather than the listed 2,500 images. This slightly overstates the digital storage requirements, making it more conservative. This estimate must be adjusted slightly if used conversely, because it overstates the capacity of the microfilm, making slightly over-optimistic for estimating the number of rolls required for a project.

ENGINEERING DRAWINGS

When folded, blueprints fit in a file drawer and have a thickness equal to the number of letter size documents that would cover the blueprints. For an E size drawing, this is 16 letter size documents because a folded E size drawing has 16 layers of paper. Using this relationship, the 50 thousand byte estimate (for a scanned page) can be used to estimate that an E size drawing would require 800 thousand bytes (16 times 50 thousand bytes) of storage. This is what is shown in the table in the pullout section that follows below.

MICROFORMS

Aperture cards are the microform most commonly used to store blueprints. Aperture cards are punch cards that have a hole (or aperture) cut into them that holds one 35 mm slide which reproduces one blueprint sheet in most cases. An image scanned from a blueprint's image in an aperture card requires the same amount of storage needed for an image scanned from the original full size blueprint.

DIGITIZED MULTIMEDIA FORMATS

Audio, Video, and Color Photographs

Multimedia documents exist in compressed digital form. The table in the pullout section that follows below shows average sizes for these documents as well. The DVD (commonly Digital Video Disc) multimedia format standards will provide a stable foundation for working with these types of documents.

Current DVD developments are posted at http://www.VideoDiscovery.com/vdyweb/dvd/ dvdfaq.html by Jim Taylor, who wrote the book: DVD Demystified: The Guidebook for DVD-Video and DVD-ROM.

PAGES PER SECOND (COMMUNICATIONS)

Communications Line Designations

There are several common communications line types available. The speed of each line type in bits per second and page images per second is given along with a rough estimate of the monthly cost of a local connection (two or three miles). This will help in confirming that the speed of access desired is possible over the communications lines proposed.

- - - Editors' note: end of part 1 - - -

OFFICE COLOR (COLOR AND GRAYSCALE SCANNING)

Office Quality Color

Office quality color generally is text or graphics without photographs. It is not necessary to exactly match colors. Red should be distinguishable from orange and from pink, but two shades of pink or a red and a red-orange color combination would not be easily distinguished on an office-quality color document.

Lossy Compression

Lossy (destructive) compression reduces the size of the image file by reducing resolution and color fidelity wherever possible while still providing the minimum acceptable image quality for office-color.

Lossless (nondestructive, isomorphic) compression, such as the CCITT G4 (Group 4) fax compression used in most document imaging TIFF (Tagged Image File Format) images, exactly reconstructs the compressed image, bit by bit, pixel by pixel, during decompression. A decompressed image is identical (isomorphic) to the original image before compression.

Lossy Compression and Records Management Procedures

The use of lossless compression may give the false impression that an image is unaltered. In fact, document enhancement techniques such as dynamic thresholding, deskewing, and despeckling, all modify the raw scan data of the scanned image to improve OCR results and improve the results (compression ratio) of lossless compression that is subsequently applied. These document enhancement techniques are the digital equivalent of the ways in which xerographic copying or microfilming modifies a document.

It is probably better to focus on the records management issues of standard-business-practices and chain-of-custody rather than alteration of the raw scanned image. For example, one would want all the records in a given records series (all records of the same type, such as invoices) to be digitally processed in the same way (including the setting for loss-level), ideally without human interaction, using the same imaging processing algorithms.

One would strive to avoid individually processing of any scanned records (modifying the processing procedure for manually selected records). If there is human interaction, it must be done within written guidelines, with no influence from those persons who are responsible for the content of the documents.

After the images of the records are processed (using an automated method), the rule of unalterability within the chain-of-custody would be followed.

Color, how much is enough?

Grayscale and color scanning can be divided into 5 quality bands. The first is view-only-color, the second is grayscale-OCR-color, the third is bi-tonal-OCR-color, the fourth is visually-unaltered-color, and the fifth is raw-color.

The spelling of Grayscale

Gray can be spelled 'gray' or 'grey'. The Microsoft Word spellchecker prefers 'gray' to 'grey' and 'grayscale' seems to predominate over 'greyscale'.

View-Only-Color 100 dpi, Lossy Compression

Most faxed pages have a resolution of about 100 x 100 pixels (dots) per inch. If high resolution is selected, the fax images are about 100 x 200 pixels (dots) per inch.

Office quality color at 100 dpi is adequate for many document imaging applications. It is not adequate for OCR (Optical Character Recognition). If the original documents are destroyed after scanning in office quality color at 100 dpi, then OCR can never by used to search the documents.

An average view-only-color image is about 100 KiloBytes.

Grayscale-OCR-Color

Grayscale-OCR is unusual in document imaging; most document-imaging OCR is bi-tonal (black and white).

At lower resolutions (below 300 dpi), in bi-tonal images, the pixels are too big to fit between the letters of a word, and adjacent letters appear to be connected. If the pixels were smaller (and the dpi higher) then a row of white pixels would separate the two groups of black pixels that form two adjacent letters. Because the letters appear to touch at scan resolutions below 300 dpi, bi-tonal OCR works poorly, or not at all, at scan resolutions below 300 dpi.

In grayscale-OCR, the character recognition algorithms (programs) are able to make use of the fact that pixels on the edge of characters appear gray rather than black because they include some of the white background around the character. These gray pixels show the separation between adjacent letters.

Grayscale OCR becomes workable at about 150 dpi. Increasing the dpi improves the results of grayscale-OCR up to about 300 dpi where bi-tonal OCR becomes feasible.

An average grayscale-OCR-color image is about 500 KiloBytes. (500 KiloBytes is very conservative. For a specific application the compressed image size may be less than 350 KiloBytes.) An estimate of 500 KiloBytes also allows for the possible introduction of 200 and 240 dpi scanned color images (Current office-color images are scanned at 150 dpi for grayscale-OCR.).

An estimate of 500 KiloBytes for grayscale-OCR-color is also conservative enough to allow for the inclusion of a 50 KiloByte bi-tonal image, in the 500 KiloByte per image storage budget, to provide lossless compression and bi-tonal-OCR.

Because the bi-tonal image is compressed losslessly, and because there is considerable experience with (and acceptance of) such images, the meaning of hybrid systems may shift from being a combination of microfilm and digital imaging to being a combination of losslessly compressed bi-tonal images and some form of lossily compressed office-color images.

3D OCR

Grayscale OCR is also called 3D (three-dimensional) OCR because it adds the dimension of color depth to the height and width dimensions of a scanned image.

Bi-tonal-OCR-Color

Bi-tonal-OCR does not work well (and in some cases does not work at all) at resolutions below 300 dpi

Bi-tonal OCR improves, as resolution is increased, up to about 600 dpi, depending on the size of characters in the OCRed document image.

No amount of processing can increase the dpi of an image. Also, no amount of processing can recover grayscale or color once an image is thresholded to a bi-tonal image. (Similarly, once an image is converted from color to grayscale the color information cannot be recovered – notwithstanding colorized movies.) (Colorized movies, and digitally restored or updated movies, are a well known example of 24 bit color scanning, and printing, of microforms – the movie film.)

Bi-tonal OCR will not work on a 150 dpi image, and grayscale OCR will not work on a 150 dpi image that has been thresholded to one bit. (Thresholding to 1 bit reduces the color-depth to 1 bit per pixel and makes the image a bi-tonal image.)

A bi-tonal-OCR color image would be about 1 MegaByte when lossily compressed. (1 MegaByte is very conservative. For a specific application the compressed image size may be less than one-half MegaByte.)

Because the compressed bi-tonal-OCR image size is so large, it may be better to save a 300 dpi bi-tonal image (50 KiloBytes) and a grayscale-OCR-color image (450 KiloBytes) to reduce storage requirements. In this case, bi-tonal OCR-color images may never be used in practice; they will always be replaced by the combination of a bi-tonal image and a grayscale-OCR-color image.

Visually-Unaltered-Color

A visually unaltered image is an image that does not appear to have been altered by lossy compression. This is the quality of color found on a Kodak Photo CD.

Because office quality color deals with graphic documents rather than continuous tone photographs; the file size of a visually-unaltered office quality color document (page) will be smaller than the file size a corresponding visually unaltered photograph. A Kodak Photo CD 16-base image (photograph), blown up to fit within an 8 by 10 inch size frame would correspond to a 300 dpi letter size image.

To achieve visually-unaltered-color, a Kodak PhotoCD uses lossy compression settings that yield an average compressed image size of about 5 MegaBytes per 300 dpi photographic image (a 16-base image blown up to 8 by 10 inches). Office (business) color would use about 2 MegaBytes per page image. This estimated storage figure could vary widely depending on the composition of the scanned documents and the degree of conservatism in the setting of the degree of loss setting in the compression algorithm.

Raw-Color

A raw-color image is the image before image enhancement and compression.

If raw images are stored, a suggested enhancement algorithm should also be identified so that the intended appearance of the image can be reproduced for users viewing the image.

Raw images may be reduced in size by as much as 20 to 30 percent using lossless compression. Because the reduction in size is less than 2 to 1, many system designers leave the images uncompressed (even though the compression is lossless) because it is easier for system users, administrators, and owners to understand the implications of unaltered images than to understand the implications of losslessly (isomorphically) compressed images (even though, mathematically, there is no difference between a raw image file and a losslessly compressed file produced from the raw image file.)

A raw 300 dpi image has 90 thousand pixels per square inch (per 625 square mm) (300 dpi x 300 dpi = 90 thousand dpsi (dots per square inch)). There are about 100 square inches (62,500 square mm, about .05 square m) in a letter size page, yielding about 9 million pixels per page. This can be rounded to 10 million pixels per page to allow for overscanning of the image to make certain that all of the edge of the document is captured (not accidentally cropped off because of slight misalignment or misregistration of the document during scanning). For 8 bit images, this would be 10 MegaBytes per page.

A 600 dpi image would be four times as large (twice as many pixels in both heights

and width), or about 40 million pixels (40 MegaBytes for 8 bit grayscale and 120 MegaBytes for 24 bit color.

Engineering Drawings on Aperture Cards

View-only-color quality, at 100 dpi may be adequate for engineering drawings because OCR is rarely used on engineering drawings and if the lettering on the drawings was done mechanically (meets drafting specifications) then 100 dpi can easily resolve the letters. 100 dpi may not be adequate for raster to vector conversion of drawings. 150 dpi grayscale-OCR-color also may not be adequate for raster to vector conversion of drawings.

Aperture cards scanners are generally grayscale or bi-tonal. There are very few color aperture card scanners. However, many of the office-color compression algorithms treat grayscale images as color images. This is why the estimates for compressed engineering drawings are based on compressed office color. As more statistics on compressed grayscale engineering drawings becomes available, these estimated compressed file sizes may be reduced.

High end aperture card scanners scan a microfilmed E-size (3 by 4 feet) (A0: 1 square meter) image at a resolution of 400 dpi, with an overscan area large enough to scan the taped area between the edge of the drawing and the edge of the aperture (or hole) in the aperture card (punch card). This produces an image of approximately 20 thousand by 16 thousand pixels or 320 million pixels (320 megapels). The raw image size is 320 million bytes because almost all aperture cards are only black and white, so 24 bit color scanning is not done, only 8 bit grayscale scanning. This file size would require one single sided DVD to store 10 scanned drawings. One DVD per 10 drawings, while very useful for samples, would be too cumbersome in most applications.

Thumbnails

Thumbnail images are used to provide users with a quick scan of a document. The art of book design provides the underlying technology on which the efficacy of thumbnail images depends. Documents are designed to be flipped through quickly to locate a section of interest. This flipping competes with the table of contents, the index, and full text indexing for locating the desired location in a document. Book design provides for all chapters to begin on odd numbered pages in a stylized format that includes a large amount of white-space and the chapter name in large print. Similarly charts and tables are easy to view at high speed (flipping) and as a result, at a low resolution.

Most thumbnails are about 100 by 100 pixels. Kodak PhotoCD thumbnails (1 / 16 base) are 128 x 192 pixels in size.

Document-imaging thumbnail images appear about 1 inch high on a 100 dpi screen (19 inch viewable diagonal, 1600 x 1200 resolution).

Variances

For office color images, the amount of loss is user selectable in most types of lossy compression by using a loss-level setting. Therefore, the variance between actual average digital image size and the estimated file sizes includes the effect of the loss-level setting selected.

The method of correcting for variances is still the same, however. The average image size is calculated for a representative sample of actual scanned images. The variance is then calculated. The variance is used to calculate a correction factor and the correction factor is applied to the estimated system capacity requirements.

- - - Editors' note: end of part 21 - - -

MORE ON COLOR (MUCH MORE) OPTIONAL SECTION

Are dots pixels?

Yes.

In document imaging, dots are pixels.

Halftone Dots

In offset printing, halftone dots are used to represent grayscale and color. The can be seen by looking at any printed picture (halftoned image) with a magnifier. As can be seen with a magnifier, halftone dots are evenly spaced in a two dimensional grid, just as document imaging dots are evenly spaced, in a two dimensional grid. Unlike document imaging dots, halftone dots vary in size (diameter). Larger halftone dots appear darker in the halftoned image; smaller halftone dots appear lighter in the halftoned image.

Halftone dots are different than document imaging dots. Each halftone dots is a macropel array of 16 by 16 pixels used to represent any one of 256 shades of gray by printing from 0 to 256 of the pixels as black, starting with one black pixel in the center of the macropel array. Additional pixels (adjacent to the first black pixel or group of black pixels in the center of the macropel) are added to the macropel array to achieve any desired shade of gray.

As the halftone dot at the center of the macropel grows with the addition of black pixels, the shade of gray represented by the growing halftone dot becomes darker.

Dithering

If the black dots added to the macropel are scattered around the macropel rather then being concentrated at the center of the array, the effect is called dithering.

On low resolution laser printers (300 and 600 dpi) the halftone dots (Each halftone dot is created using a 16 x 16 pixel macropel array.) are so large that the eye becomes focused on the dots and does not see the shades of gray that are intended to make up the continuous tone image. Dithering breaks up the individual halftone dots so that they do not attract attention. (Technically, dithering increases the areal frequency of the image so that the areal frequency of the image exceeds the areal bandwidth of the eye (exceeds the areal density of rods and cones), causing one's visual system to blur the image, creating the illusion of shades of gray.)

Unfortunately, the resolution (screen ruling) of halftone dots is just inside the capabilities of offset printing. Common screen rulings for halftone dots are 65-85 dpi for newspapers, 133-150 for magazines and books, and 175-200 for high quality magazines, high quality books, and for magazine covers. Dithering is at the resolution of laser printers, usually 300 to 600 dpi. As a result, an offset (halftone) printing press (or a mid- to low-end copier) cannot reproduce the dithered image.

To correct the problem, the dithered image must be mechanically screened, just as though it was a photograph, to produce halftone dots the printing press can reproduce. For copying, high-end copiers include a halftone screening mechanism that improves their ability to reproduce photographs (and dithered images).

Black and White vs. Grayscale

Black and white television actually is grayscale television.

Black and white scanning assigns one of only two values to each pixel, either a '1' or a '0', and the pixel is either 'black' or 'white'. Because only 1 bit is needed to represent a value of '1' or '0', black and scanning is often referred to as one-bit scanning. Black and white scanning is also called bi-tonal scanning because it only uses two (bi-) tones (black and white).

1 Bit, 2 Grayscale Values

Strictly speaking, the two tones (shades) of black and white scanning constitute a grayscale with two values: black (0) and white (1).

2 Bits, 4 Grayscale Values

If a second bit is used to divide white (1x) into white (11) and off-white (10), and to divide black (0x) into gray (01) and black (00) then a grayscale of four shades can be represented by the two bits.

3 Bits, 8 Grayscale Values

Each of the four shades (in 2 bit grayscale) can be divided into a lighter and darker shade (producing 8 shades) with the addition of one more bit. White (11x) is divided into light-white (111) and dark-white (110); off-white (10x) is divided into light-off-white (101) and dark-off-white (100); gray (01x) is divided into light-gray (011), dark-gray (010), and black (00x) is divided into light-light-black (001) and dark-black (000).

8 Bits, 256 Grayscale Values

Adding a 4th bit to the 3 bits used to represent 8 grayscale values allows to division of each of the 8 value or shade of color into two shades, a lighter one, and a darker one, producing a total of 8 x 2 or 16 shades.

Adding a 5th bit produces 16 x 2 or 32 shades. A 6th bit produces 32 x 2 or 64 shades. A 7th bit produces 64 x 2 or 128 shades. An 8th bit produces 128 x 2 or the familiar 256 shades of gray, also know as 8 bit color. Eight bit color is very popular because it uses exactly one byte for each pixel.

Color Depth

Color (including grayscale) is given as the number of bits per pixel (image depth), or the number of shades of gray recorded per pixel. As seen above, the number of shades of gray maps directly onto the number of bits per pixel. More precisely, the number of shades of gray per pixel is defined by the number of bits per pixel.

Grayscale, a Subset of Color

Color can be expressed as the sum of three grayscale values, one value for Red, one for Green and one for Blue (RGB). These values can also be express as (and transformed into) Hue, Saturation, and Intensity, or as CMYK color, (Cyan, Magenta, Yellow, and K for Black)

Assigning 8 bits to each of three colors uses 3 x 8 bits or 24 bits to represent a color pixel. This is the origin of the term 24-bit-color. Grayscale is a special case of color in which each pixel has the same value for each of the three colors. For example, white is represented as (255, 255, 255) (red = 255, green = 255, blue = 255) and black is represented as (0, 0, 0) (red = 0, green = 0, blue = 0)

24 Bit True Color

As noted above, 8 bits can represent 256 shades of gray. 24 bits can represent 16,777,216 colors (8 bits for the red shades, 8 bits for the green shades, and 8 bits for the blue shades that are mixed together to form the exact or true color that is recorded in the 24 bit image). For each of the 256 shades of red, any one of the 256 shades of green can be added creating 256 shades of red-green. This produces 65,536 shades of red-green. Each of these 65,536 shades of red-green can be mixed with the 256 shades of blue creating 16,777,216 shades of red-green-blue.

Continuous Tone Images

24-bit-color images are said to be contone (continuous tone) images (or true color) because the blending of colors appears to be continuous between pixels. If fewer color (bits per pixel) are used, then there are sharp breaks between different colors (With fewer available colors for rendering an image, the difference between adjacent or similar colors is correspondingly larger.) creating a posterized effect (posterization).

24-bit color is also called true color because it can exactly match (is visually indistinguishable from) the intended color. This is important in advertising because customers come in to buy the color of the material they saw in the ad. The ad color must exactly match the color of the actual material, or the customer will be disappointed. Also, readers who are looking for the actual color will not come in because the actual color did not appear in the ad. Office color is not concerned with true color.

Graphics

Graphics images are images that contain black images (characters, symbols, trademarks, or drawn shapes) on a white background. Any single color image on a single color background is also considered a graphic image.

Tints

A uniform background color is often called a tint which is a halftone pattern using a single size of halftone dot which produces a uniform background of a single color or shade.

Frequently, to make forgery detection easier, a continuous tone tint us used in business color. The tint varies from dark to light, from the top to the bottom of a document or vice versa. The tint variation sometimes occurs several times, in multiple bands, from the top to the bottom of the document. This background continuous tone tint is extremely sensitive to erasure, and clearly shows any attempt to erase characters printed on a document.

These continuous tone tints increase the size of compressed business quality documents. However, because the variation on the continuous tone tint is extremely uniform, the tint compresses much better than a continuous tone photographic image.

Cartoons

Comic strips and comic books use black lines to outline the characters and objects in scenes. Color is then splashed in, often not in close registration (running over the black lines, or not completely filling in between the black lines). This technique of black outlines and color splashes produces quite acceptable images because our visual system processes color in three separate and distinct ways: resolution, color, and motion.

Resolution (and symbols) are processed in black-and-white and color is processed for color depth and recognition, but not resolution. This separate processing is why the black lines in coloring books help to improve the appearance of colored images even though the children doing the coloring may go outside the lines. (The equivalent of the enhancing effect of the black lines in coloring books is that technically, in lossy color compression, the chrominance is compressed more than the luminance. In the future the compression algorithms can be further biased toward resolution (luminance)

allowing higher resolution scans to improve OCR while preserving the lower resolution chrominance needed for image recognition by users.)

The separate paths of image processing, for resolution and color, are also the origin of the expression 'operating outside the box' (or lines), popular in management training.

The third image processing track is motion. Our ability to detect motion is most sensitive at the periphery of our vision (Presumably this is where the tigers were when our visual system was evolving.). This is why we see video monitor screen flicker on a monitor on a desk adjacent to our own more easily than on our own monitor. Florescent flicker is also more easily seen out of the corner of one's eye.

Manipulating the three image processing paths independently results in image processing errors in the human visual system. These artificially, and usually purposefully, induced image processing errors are commonly known as optical illusion.

- - - Editors' note: end of part 3 - - -

DVD AND CD

The CD

The CD (Compact Disc) development was funded by the music industry. A CD can hold about 650 MegaBytes of storage. In this paper, a CD is conservatively estimated to hold 500 MegaBytes in actual practice. This is consistent with the goal of erring on the conservative side and slightly overestimating the storage required at every step. With these estimates, one CD can exactly hold the scanned contents of one standard file cabinet, meeting the goal of producing simple, round number, estimates.

The DVD

The DVD (commonly Digital Video Disc) has been funded by the video industry, the movie industry, the music industry, and the computer industry. The DVD is about to unify the PC, TV, telephony, and document management. The DVD has several versions and options. The table entry shows the capacity of each version. Also shown are the DVD format specification for recording video and audio.

The estimate of 10 file cabinets per DVD is very conservative. It assumes considerable overhead for storing indexes, and gives a large amount of weight to the goal of creating round number estimates, using a figure of 10 rather than 12 file cabinets per DVD.

Spelling 'Disk' and 'Disc'

For magnetic disks, use 'disk' with a 'k'. 'Disk' follows the metal disks in a harrow. For optical discs such as DVD and CD use 'disc' with a 'c'. 'Disc' follows the spelling of music record discs. When referring to disks and discs collectively, 'disk' is used in this paper.

PIXELS, HOW MANY?

Pixel Sizes, Pixels per Image, and . . . Sizeless Pixels

Pixel size is based on the image or object scanned or illustrated (or the size of a printed or displayed image), for example 1 / 300 inch (square) pixels (~~.1 mm or 100 micrometer square pixels) scanned from a letter size page scanned at 300 dpi. (~12 or ~~10 dpmm) If the image has been microfilmed at a reduction of 12X then the pixels scanned on the microfilm are scanned at 3600 dpi (1 / 3600 inch square) (~~.01 mm square = 10 micrometers square) to have an effective resolution of 300 dpi (~12 dpmm) relative to the original letter size page. If the pixels are displayed on a monitor that has a resolution of 100 dpi (4 dpmm), then the pixels are 100 dpi (1 / 100 inch square) (~.25 mm or 250 micrometers square), and have been enlarged, along with the document, by a factor of 300 percent (3 times) from the original document. If the pixels are combined, 9 pixels to 1 pixel (3 pixels to 1 pixel in both dimensions of the two dimensional image), to display the letter size document at a 1 to 1, normal size, on a 100 dpi monitor, then the pixel size in increased and the image resolution is decreased.

When speaking of pixels per image, the number of pixels, and the amount of information in the pixels stays fixed, no matter what size the image is reproduced. This relationship is most clearly seen in a store selling many sizes of television sets. If the same demonstration video clip is being shown on every television set, then every set has the same picture, with all the same picture information, no matter what size the television set is.

Pixels do not have a size in the computer or when they are stored digitally. A raster of digitally stored pixels is just an array of numbers. This array of number carries all the image information.

Meta-data is the information carried with the raster (array) of sizeless pixels (that make up an image) that includes the information on the size of the pixels in the scanned (or original) object, the size of the pixels in a scanned micro- or macro-form of the object, and the size(s) of the pixels in an intended reproduction or reproductions of the object or document.

The digitizing process

Scanning a document creates a digital image that is an analog (physically similar to) of the paper image.

UNITS OF MEASURE (DIGITAL)

Commercial vs. Computer Size of a KiloByte, Kilobit (etc.)

Why is a computer MegaByte not exactly one million bytes? Because computers use binary arithmetic and the closest round number computers have that is near one million is two to the twentieth power. Two to the twentieth power (2**20) (When writing exponents, it is best to express the values in such a way that editors and proofreaders not familiar with exponents will be able to accurately reproduce the meaning of the expression.) is equal to 1,048,576. This is why many computer program displays show an exact number of bytes beside a seemingly smaller number of MegaBytes. (For example, the Disk or file properties window in Windows 95). Similarly, a computer KiloByte is not exactly one thousand bytes, but is two to the tenth power (2**10) or 1,024 bytes.

These differences are so small that they do not belong in a management discussion, but lawsuits have been brought over them. Being kind and understanding to those who bring them up is the fastest way to move on to a productive discussion.

Because of the lawsuits over the meaning of KiloByte, MegaByte, etc., only the metric meanings (based on units of 1 thousand) can be used in commercial discussions of storage capacities. The computer-based terminology, based on units of 1,024 = 2 to the 10th power, will continue to be used in discussing computer configurations because computers actually use equipment based on units of 1,024. The computer units will always be converted into commercial metric (1 thousand based) units before commercial discussions take place. Document imaging and document management discussions fall into the category of commercial discussions.

PAPER, TREES, COLD, and SCANNING

Trees

Just a note on the environmental aspects of these figures: Each full file cabinet represents one pulp tree. Pulp trees are grown fast for paper or are trees culled from among the trees that will be allowed to grow to a larger size for lumber.

Word Processor Pages and COLD Pages

When documents are imported directly in the form they were created in, they require much less storage space. Because COLD (Computer Output to Laser Disc) pages come from mainframe computers, their formatting is very simple and requires even less storage space than word processor pages.

Index Storage Requirements and OCR Pages

Storage requirement estimates can be done more quickly by ignoring the storage requirements for OCR images and indices. This shortcut will not greatly affect the accuracy of storage estimates because of the small size of text and indices relative to the size of compressed scanned images. If some of the images will be on optical media, but all of the indices and OCR text will be on magnetic media, then the additional accuracy provided by the following section may be useful.

OCR (Optical Character Recognition) for full text storage produces the largest indices. At 5 thousand bytes per 50 thousand byte raster scanned page image, the OCR text takes up ten percent of the storage in a system. An additional 2 to 10 thousand bytes of are required for the actual index to the OCR text, the full text index, depending on the indexing software used.

Key word and database indices rarely have more than one hundred characters per document. For one page documents, 500 bytes is one percent of the document page image size, so the database entry is less than one-fifth of one percent of the page image size, and can be ignored for most systems.

As with scanned images, the size computer generated text files can be very accurately estimated by measuring the first one percent of the documents processed when the system goes into operation.

Five Cents (US Dollars) per Page for Scanning

Storage costs are very often less than ten percent of the cost of document conversion. This estimate will provide a basis for assigning the best relative weights and values to the cost of storage and scanning in evaluating system feasibility.

CONCLUSION

Summary: Ignition, Dialog

These estimates are intended to assist managers in creating a dialog with all parties involved in a document imaging system. The 'I agree with that.' sometimes means 'I did not read it.' The best dialogs start with 'That's wrong, here is what it should say, and this is why it should say it.'. During implementation, there is nothing better than: 'My estimate says X and your report say Y, why are they different?'

From the records manager who say 'You missed two of the records storage rooms.', to the systems staff member who says 'We did not know you needed 100 Megabit ethernet.', this paper is intended to help everyone involved in a document management project contribute to its success.

This paper refers to the tables in white paper 22009, Computer Storage Requirements for Various digitized Document Types.

- - - Editors' note: end of part 4 - - -

Note to Readers

Updates and More Detailed Descriptions

When using the information in this article, please check the website http://www.ArchiveBuilders.com for updates. The version number of this article is just before the page number below. The website also has articles that provide more details on some of the terms and concepts in this article.

Comments

Please let us know how you like this paper, or if you had any questions. What would you like to see in the future? Also, please let us know where you saw this paper. For more, and the most recent version of this article, please visit our web site at http://www.ArchiveBuilders.com

Please send your comments via email to SteveGilheany@ArchiveBuilders.com. Tel: +1 (310) 937-7000 Fax: +1 (310) 937-7001. Also, please let us know where you saw this article.

Acknowledgements

Reprinted from Archive Planning, Volume 5, number 3, 2001, Archive Builders' analysis newsletter for document management.

See http://www.ArchiveBuilders.com

Note to Editors

Paper 22003v035

We will continue to update these articles as we get comments. Please contact us for the most current version before you publish. Also, please request permission to publish the article. Permission will be given freely for most purposes.

Steve Gilheany
Archive Builders
1209 Manhattan Ave., C-14
Manhattan Beach, CA 90266
Tel: +1 (310) 937-7000 Fax: +1 (310) 937-7001
SteveGilheany@ArchiveBuilders.com

Dividing this Article into Parts for Serialization

This paper is divided into 4 parts: part 4 describes page 1 of the pull out section, parts 2 and 3 describe the first part of page 2 of the pull out section, part 4 describes the remainder of the pull out section.

If you decide to divide this paper into parts please print at least the updates, comments, and acknowledgements sections in each of the parts along with:

'by SteveGilheany@ArchiveBuilders.com'.

Bio

Steve Gilheany, BA in Computer Science, MBA, MLS Specialization in Information Science, CDIA (Certified Document Imaging System Architect), AIIM Master (MIT), and AIIM Laureate (LIT), of Information Technologies, CRM (Certified Records Manager, ARMA) has eighteen years experience in document imaging and is a Sr. Systems Engineer at Archive Builders.

Author

Steve Gilheany is a Sr. Systems Engineer at Archive Builders. He has worked in digital document management and document imaging for twenty years.

His experience in the application of document management and document imaging in industry includes: aerospace, banking, manufacturing, natural resources, petroleum refining, transportation, energy, federal, state, and local government, civil engineering, utilities, entertainment, commercial records centers, archives, non-profit development, education, and administrative, engineering, production, legal, and medical records management. At the same time, he has worked in product management for hypertext, for windows based user interface systems, for computer displays, for engineering drawing, letter size, microform, and color scanning, and for xerographic, photographic, newspaper, engineering drawing, and color printing.

In addition, he has nine years of experience in data center operations and database and computer communications systems design, programming, testing, and software configuration management. He has an MLS Specialization in Information Science and an MBA with a concentration in Computer and Information Systems from UCLA, a California Adult Education teaching credential, and a BA in Computer Science from the University of Wisconsin at Madison. His industry certifications include: the CDIA (Certified Document Imaging System Architect) and the AIIM Master (MIT), and AIIM Laureate (LIT), of Information Technologies (from AIIM International, the Association of Information and Image Management, www.AIIM.org), and the CRM (Certified Records Manager) (from the ICRM, the Institute of Certified Records Managers, an affiliate of ARMA International, the Association of Records Managers and Administrators, www.ARMA.org).

Contact:

SteveGilheany@ArchiveBuilders.com
Tel: +1 (310) 937-7000 Fax: +1 (310) 937-7001

For more information, courses, and papers:

http://www.ArchiveBuilders.com