Document Scanning Considerations

How are your Documents Stored:

- How many boxes of paper do you have? If they are not in boxes, how many file drawers or how many linear feet of shelving does the collection span? (13" of shelving will fit in a bankers box with a little room and a file drawing typically spans 1.5 bankers boxes)
- What size are the boxes, are they a standard bankers box 15" long or the double size at 24" (A standard bankers box typically yields 2300 images, where a double is closer to 5000 images. The number of scanned images will depend on how many folders are present and how tightly packed are the boxes.)
- Do you have a mix of single and double sided pages?
- Is the content the same across the boxes?
- How is the content grouped and sorted? eg. Financial documents Grouped by year, in folders by payee then relatively sorted by date. Or Patient Files grouped by year, then sorted by last name.
- How many folders per inch or per box? (this allows us to estimate how much indexing is required. If it varies, take 5 minutes to count the folders in two boxes or in 24in and average)
- How are the pages stored in the boxes, are they loose in folders, in three ring binders, stapled on two prong top posts, comb or spiral bound.
- Are the documents mostly 8.5x11, do you have small receipts, 11x17, large drawings, greenbar accordion type perforated pages, sticky notes, etc?

Indexing: How do you find a document currently?

- If you can describe the process you use to retrieve a document manually that will help the vendor understand how to index the files so this can be mimicked electronically.
- What level of indexing do you require? Do you want one pdf file for the entire box with a range (eg. Smith_through_Wilson.pdf), one file per folder (eg. Smith_Joan_P.pdf, Smith_John_W.pdf…) or one pdf for each document type (eg.Smith~John~P_Application.pdf, Smith~John~P_Labs.pdf , Smith~John~P_Corespondance.pdf) or one pdf for each document (eg. Smith~John~P_email 20070315.pdf, Smith~John~P_Memo 20100224.pdf, Smith~John~P_Insurance.pdf, Smith~John~P_InsuranceRevisionA.pdf)
- The level of indexing is important to understand as the more indexing provided the quicker and easier you can get to a specific document without searching, however someone has to type all of that information and it can drive the conversion costs. However if it is only indexed at the box level (which is very rare) you may have to page down through 2300 pages to find what you

need. The balance is typically determined by how often you expect to be in the boxes. If it is a pure archival requirement for inactive customers/patients it is common to just index to the folder level. If these are active customer files which you will be importing into a document management system and adding files plus searching the old content frequently, it may make sense to index down to document type or every document.

- Does each folder have an index value which is eye readable, like a patient name, client name, employee number, date or description? Is this information also on the first page of the file or will the vendor need to page through the file looking for the index information?
- Do you have an electronic list of the index information? Often this is available in an accounting or scheduling system and can reduce the cost of the conversion effort and provide very high data accuracy.

Document Preparation for Scanning

This can be one of the biggest unknowns to a customer and a vendor pertaining to how long and how much cost to estimate for document prep. Even with 20 years of experience with many hundreds of millions of staples pulled, no vendor can precisely estimate exactly how much time it will take to prep your archives, nor do we expect our customers to know any better. What is most common is to ask some basic questions about how the originals are bound and organized and estimate how long it will take per box.

What is paramount is to test your vendor on the first two boxes and have them time how long it takes to prep, then have them scan and bill you for the first boxes. This allows you to review their work and you have a baseline to go with the estimate. This avoids the dreaded and unfortunately all too common final bill that is twice the price of the original estimate.

- How heavily stapled are your documents. Use the following as a guide:
    - Light Prep: Loose pages in folders, a staple every so often with documents 20+ pages. ~100 staples per box. Paper clips common rather than staples and items in 3 ring binders, comb and spiral bound documents 100+ pages
    - Medium Prep: Items hole punched at the top on two prong posts with multiple sections and progs per folder, 5 page documents stapled together. Occasional receipts, sticky notes and larger format like 11x17
    - Heavy Prep: 2 page documents, receipts stapled, multiple documents stapled together where 2 or more staples need to be pulled to release the pages. Receipts loose and floating around or taped to originals, frequent sticky notes. Pages still in envelopes, perforated pages (like the old green bar tractor feed stuff)
- Do you require that the originals be deprepped? Deprep is the process of replacing the binding and staples. If deprep is required it complicates the entire prep and scanning process. Either we need to insert tags to identify where documents are bound or we have to prep at the

scanner. This adds time and slows the scanning process. If you are considering deprep it is wise to get a quote with and without deprep so you can understand the cost impact. 99% of scanning is performed to either destroy or move the originals to long term storage where they are almost never accessed. In these solutions we have strategies to tag the document breaks without rebinding as an option.

- What should vendors do with sticky notes? Most often they are moved to an area of the page where it does not obscure data and are taped down or they are taped to the back of the page.

Scanning Methodology

- What resolution do you want or need to scan at? There is a lot of misinformation in the industry about what is an acceptable resolution to scan at. Washington State Archives "recommends" 300dpi uncompressed scanning and many individuals take this as a requirement. It is certainly not a requirement and it is a resolution which is more than the industry standard of 200dpi scanning. The industry uses 200dpi for many reasons.
  1. Higher resolution does not mean better legibility. Higher resolution is beneficial for capturing fine detail, higher bit depth (grayscale) is for capturing faint originals and light signatures.
  2. Documents with 8pt and large text are fully legible at 200dpi.
  3. 200dpi grayscale is far more legible than even 400 dpi monochrome
  4. All of the scanners made for the last 30 years are optimized for 200dpi so there is a higher and often unnecessary cost for scanning at 300dpi or higher
  5. Engineering drawings should always be scanned at a minimum of 300dpi and often we recommend 400dpi
  6. Charts and spreadsheets or maps where the detail is very tiny can and should be scanned at higher resolutions.
  7. Artwork, slides or photographs, particularly if they are going to be enlarged, should be scanned at a resolution equivalent to create a 200dpi print. For instance a 5x7 photograph should be scanned at 314dpi to create a 200 dpi print at 8x10.
- Are the originals single or double sided. If you are unsure or if you are concerned about a page filed upside down you may want to have your vendor scan in duplex with automated blank page removal for the backside. This is the most economical solution and insures full capture of your collection. The automated process can retain backsides with stray marks or content that bleeds through but it is typically less than 1% and is much less expensive than manual removal of backsides
- Do you have color content, light date stamps or signatures, faded faxes. The scanners scan at the same rate in monochrome or grayscale and they have the ability for auto-color recognition. Most vendors over the years have used monochrome scanning because the scanners were 5x as fast in monochrome and because the file sizes are 5x smaller. This is no longer the case, the

scanners are fast in grayscale and the cost of storage means the file size is not as much of a concern. A monochrome scan can have the effect you notice when you try to photocopy your driver license and picture goes black. You have to put the copier in photo/gray mode to get a legible copy. The same is true for scanning in monochrome vs. grayscale.

- Is there content on the face of the folder which must be captured
- Do you need your documents rebound or are they going for destruction. Rebinding can be quite costly as the vendor not only has to spend the time rebinding, they also need to modify their scanning process so they know where each document stops and starts so they can be properly rebound.
- What should be done with sticky notes? Often these are moved to an area of the page that does not obscure data and taped down or taped to the back.
- Do you need additional indexing and is any of it already available electronically? For instance do you want to search the collection by first and last name, SSN, Date Of Birth, etc. Do these things need to be manually keyed or can we key the SSN then get the other information from a database or spreadsheet.

<u>Delivery of the Image and Index Data</u>

- What format do you want or can your document management system accept? PDF and PDF/A are very common, however technically it is a proprietary format. Tiff GroupIV is common for monochrome imagery and will not support grayscale or color. JPEG or JPEG2000 is common for gray and color imagery.
- How should the index information be delivered? Common formats are Excel, Access or CSV with a field which pairs the index data to the image path. Some customers will want the basic index information in the file name (eg. LastName_FirstName_DateOfBirth.pdf) while others will want all of the metadata stuffed into the header of the imagery.
- Are you just looking for images and index information on a DVD or Hard Drive or are you looking for a simple search engine or a hosted solution.
- Where will you store the information? Copied to a server, hosted or a DVD based system. You may want the vendor to store a copy on DVD. Certainly you will want to keep a copy on DVD offsite which is only accessed in a data loss situation. DVDs are typically rated for 25 years but you can lose data if they are scratched.