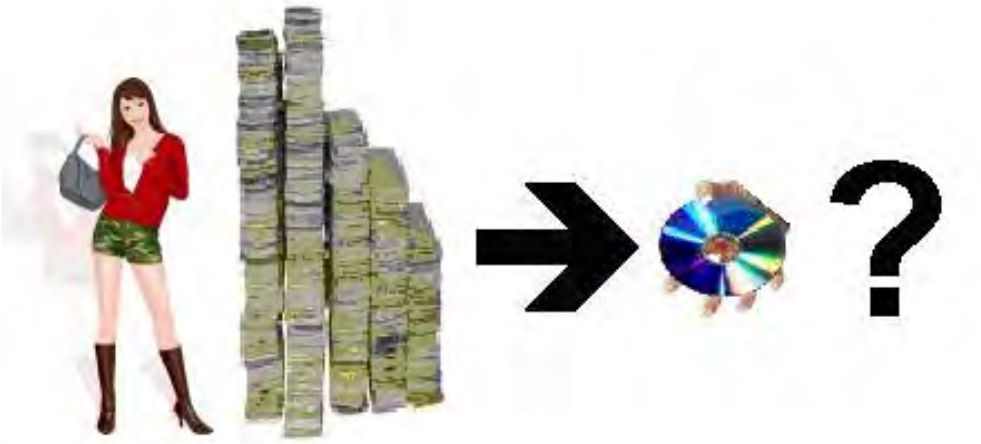
DOCUMENT SCANNING

A Brief Guide



Merlin DMS Limited
Mandora House
Margaret Louse Road
Aldershot
GU11 2PW
Tel +44 (0)1252 336363
Web http://www.merlindoc.com
Email info@merlindoc.com

© Copyright 2013 Merlin DMS Ltd

A BRIEF GUIDE TO DOCUMENT SCANNING

CONTENTS

1. Why scan at all?

- The most up-to-date and dynamic business information belonging to an organisation is held in computer systems and can usually be retrieved, processed and distributed quickly, easily and at will.
- This, however, is usually a tiny proportion of the legacy data that is held in paper documents that are held in various locations, formats and states of disorganisation.
- Access to this data is rare, and usually less than 1% is ever needed, but there is no way of knowing which and the whole lot must be kept.
- If a file is removed for any reason, it is quite often not returned with the same care that attended its original placement, which can result in that document being effectively lost.
- Worst-case scenario of a lost document varies from one organisation to another, but it can be terminal, as in the case of a well known transport company that folded under the weight of uncollected debts resulting from its inability to find PODs.
- Even if this extreme case does not apply, some alarming statistics:
    - *Office-workers spend between 30 and 40% of their time doing non-productive document related tasks* - Gartner Group survey
    - *80% of managers and directors of SMEs spend at least an hour each day looking for documents* - YouGov survey
    - *Managers spend up to 25% of their time searching for information* - Accenture survey

Conclusion: Scanning of documents to a well-constructed electronic document management system can save time and money expended in searching for documents and guarantees that none will ever be lost.

2. GDPR!

- Deadline **25th May 2018** – the old Data Protection Act is replaced by the General Data Protection Regulations.

- This is an EU law, but the UK Government has made it clear that Brexit will not affect the content or timing of the new rules, which, as the UK were major contributors, are likely to continue mainly unaltered and possibly even strengthened after Brexit.

- It has been brought about as a result of the increasing numbers of high-profile scandals involving leakage of personal data – the NHS, banks, mobile phone companies – which have underlined the need for a tougher security regime.

- There will be much tighter controls over the protection of personal identity data - organisations will be required to prove they have been pro-active in compliance and there will be draconian penalties for breaches.

- The major change will however be the fact that paper files will be covered , instead of just digital records which was the case with DPA, and without investing in large security strongrooms, it is hard to see how hard copy archives can meet the requirements of the new regulations.

  Conclusion: Whereas a good case for scanning can be made on cost and efficiency grounds, the new law will make it the best and cheapest option to be compliant.

3. How much paper have we got?

- A full 4-drawer filing cabinet, standard size, contains an average of 20,000 A4s
- A 3-foot shelf in a lateral filing cabinet contains an average of 7,000 A4s
- A standard size archive box (if such exists) holds an average of 3,000 A4s
- A lever-arch file contains and average of 300 A4s
- Rule of thumb: one linear foot of tight-packed filing contains an average of 2,500 sheets.
- All these averages are VERY ROUGH and huge variations can be observed: large numbers of thin paper will increase them considerably, and large numbers of cardboard folders will decrease them.

Conclusion: Use these guidelines only for estimating the MAGNITUDE of the problem: whatever figure you come to, expect a final outcome to be $\pm$ 50% or more.

4. Go Easy on the Staples

- You may not be scanning your documents at the moment, but there is a good chance you will at some time in the future and what goes in with a fraction-of-a-second click of the staple gun, takes time, considerable loss of temper, not to mention a good few broken nails and cut fingers, to remove.
- Numerous devices have been invented for removing the wretched things but none of them works efficiently.
- If and when you do get around to scanning (if you're not already doing so) you will find that liberal use of the staple machine in the past has created a long-winded unpopular job if you do it in-house, and an expensive one if you outsource.
- Instead of stapling them together, try keeping wads of related papers in clear plastic folders.

Conclusion: If you are not scanning now, or even planning to in the foreseeable future, a little application of foresight will provide big future savings in time and money should you choose to do so in the future.

5. To weed or not to weed?

- Many files contain duplicate and superfluous material, like With Compliments slips, hand-scribbled notes that were relevant at the time, duplicates etc.
- It is, on the surface, a waste of resources, whether in-house or out-sourced, to scan documents that are of no value.
- If it is possible to select whole files to be discarded, for instance those before a certain date, this is a relatively quick and simple exercise which will easily pay for itself.
- However, if we are talking about removing individual pages from files, the savings can easily be swamped by the cost
- Making a decision whether to keep or discard a document cannot usually be left to an office junior, so we are talking about resources whose opportunity cost may vastly exceed any benefit.
- If a program is planned in which the documents are prepared for scanning at the same time as weeding AND utilises otherwise spare time, then exercise becomes more attractive.
- We must also balance the cost of such a one-off exercise against the time wasted, if it is not done, wading through redundant images until you reach the right one, which will be an ongoing cost.
- It must not be overlooked that if a regime is in place that encourages people NOT to keep paperwork that is of no interest, it will not help the existing situation, but have a big impact on future scanning activity.

Conclusion: There can be good justification for an exercise to thin the files before they are scanned, but before committing to it, bear in mind that it will be a big job requiring the valuable time of relatively senior staff. Not something to be undertaken lightly and implementation a more stringent policy as to what is kept at point of filing would pay big future dividends.

6. Colour or Mono?

- Modern scanners allow documents to be scanned in colour at a speed not much less that in black and white, and if the scanning is outsourced, will be reflected in prices that are not that much different.
- This has major benefits, amon which can be mentioned:
  - Coloured areas on drawings, plans and maps have crucial significance which would be totally lost in mono.
  - Old index cards with have writing in coloured ink which merges into the background when reduced to greyscale.
  - Logos on documents that identify the source can be rendered unrecognisable without colour.
- And let's face it, documents look better in their original form - those of us old enough to remember black and white photography will recall the sheer pleasure of seeing photos in natural colour.
- But let's not get carried away as scanning in colour is not suitable for everything. There is obviously no advantage of scanning black and white originals in colour, but there is another major consideration: scanning in colour is usually a "lossy" process, which means resolution is lost. This can result in small print becoming unreadable.
- It is possible to scan to non-lossy formats such as TIFF and BMP, but the files are large, unwieldy and can often take an unacceptably long time to open.
- If OCR (optical character recognition) is required, then colour is definitely a no-no. Its "lossy" character makes it impossible to define borders of character shapes to the precision required for an OCR engine to interpret them with any accuracy. It will make a best guess which is sometimes right, often wrong, but in any case fails to provide an adequate result in almost all cases.

Conclusion: So there we have it - whether colour or black and white, you pays your money and takes your choice, but now that decision can be based on need and not speed or price.

7. How do we find our documents? Indexing

- This is not something to be considered as an afterthought as the very purpose of the exercise is to make it easier to find documents, and if it's not planned with this in mind, it will fail to do so.
- You will already have an index of sorts – the way that the hard copy is filed and this may be adequate and simply replicated in electronic form.
- The big difference with scanned documents is that you can apply a variety of searchable index information **(called metadata)** so if it can't be found one way – possibly because the enquirer has been quoted a wrong number - then it maybe possible to find it another way.
- An example – sales invoices filed by sequential number. In hard-copy that is it, and if the required item can't be found, it is a choice of giving up or looking through the whole filing system until it is. More information can be added to scanned files so that in the same system, a document could be found by any one of, say, the following, or indeed a combination thereof:
    - Invoice number
    - Account number
    - Invoice date
    - Net amount
- Serious consideration must be given to the frequency and urgency of document retrieval before deciding on the level of indexing as the more metadata is added, the higher the workload at the scanning stage which must be justified by time saved in later retrievals.

Conclusion: The whole point of scanning is to be able to quickly retrieve documents that could otherwise have taken a long time or, indeed, never found. The options for facilitating this are numerous, but remember there is a cost in applying them and if it exceeds the likely benefit then the purpose is rather defeated.

8. **Some indexing shortcuts**

- In the last section, we pointed out the financial drawback in applying more indexing than the likely retrieval would justify, but there are ways of reducing, even eliminating, the manual content of the job, dependant on the nature of the documents.

- For the simplest case – e.g. the sales invoices as per above - it is possible to set up the scanning software to apply sequential numbering to the file name that matches the invoice numbering. Just remember to put in Missing Document sheets for any numbers that are missing, otherwise the process will get out of synch.

- Once again, taking the example above, all index fields including account number, invoice date and net amount, already exist in the accounting system and can usually be easily extracted and linked to the invoice number by standard database techniques.

- The above example is only applicable to single page documents in strict numeric sequence. However, where a document number is in a standard position on the page and is clearly printed, we can use OCR – optical character recognition – to create the index. There are dangers on relying on this, for instance, where there is a mixture of alpha and numeric characters and 5 can be mistaken for S, 6 for G, 8 for B etc – and in those circumstances, there should be some verification process.

- More accurate than OCR are bar-codes, as used in supermarkets, which are not far short of 100% reliable. To be effective, they must be produced as part of the document production process, which of course requires pre-planning.

- Documents with variable numbers of pages require a mechanism to define where one ends and the next begins. With OCR and bar-coding, it is usually possible to define the position and content of these entities, so that where they don't meet the criteria or are completely absent, it can be assumed that the page in question is a continuation. In other cases, separator sheets or patch codes – sheets containing software-recognised markings - that are manually inserted between documents. These can often be formatted in such a way as to contain readable data as OCR- or bar-codes.

Conclusion: There are a number of options available that can reduce the labour content of the indexing process, but these are not applicable in all cases and must be chosen and implemented with care and foresight.

9. Preparation for scanning

- Modern scanners, even low-cost models, can gobble up paper at an impressive rate, but the paper put in the feed hopper must be perfectly prepared.
- Staples, which have already been mentioned, are the chief enemy of the scanner: not only will they abruptly halt the process, but can do damage to the machine itself. Quite often, it will be classed as abuse that invalidates the warranty.
- Beware folded corners that obscure text – be vigilant and unfold them before scanning as they won't be noticed until the information they are hiding is required, and then it's too late.
- Repair torn pages as small pieces can become detached and lodge in the light path obscuring that part of the image, which will probably not be noticed until later, possibly too late, as a thick black line striping vertically through all images.
- Where there are documents of varying numbers of pages, don't forget to insert separator sheets as failure to do so can mean that, in the absence of a clearly defined sequence, a document can be lost forever.
- Sometimes there will be large drawings interspersed with normal A4 text pages, which must be scanned on a different machine made for the purpose. It is sometimes important to retain the position of those drawings and a good tip is to replace each with a photocopy of its title panel.
- Whatever level of preparation you need to carry out, the labour content will almost certainly turn out to be more labour than the scanning itself, so be prepared to have as many as three people preparing to just one doing the scanning.

Conclusion: Meticulous preparation is vital before the paper gets anywhere near the scanner and is likely to be considerably more labour intensive that the scanning process itself. Skimping, however, can be ruinous, not just to your throughput, but to your equipment itself.

10. What scanner?

- There is a confusing array of equipment on the market ranging from under £100 up to over £50,000.
- The main protagonists are Kodak, Canon, Fujitsu and Panasonic and they are all good in their own way for particular jobs and no specific recommendations will be made here as new models and improvements are appearing all the time.
- The cheapest models are flat-bed machines that can take anything up to a minute, or even more, to scan an A4 side. These are not serious contenders for business use, except perhaps where there are colour photos present that need high definition.
- Automatic feed scanners are available from about £300 upwards and at the lower end, there are some very good machines that can scan up to 1000 sheets per day and two or three of these will probably suffice for the ongoing volumes in quite a sizable organisation.
- Higher up the cost-scale are faster machines that, in most cases, can take sheets up to A3 in size – quite important in a common application, purchase invoices, where there are huge variations in paper sizes and certain suppliers often stray over the limits of A4.
- Right at the top are the high-speed production scanners, that are mainly of use to service bureaux, but which are also suitable for high volume heterogeneous applications such as exist in banks, insurance companies and government offices.
- Nearly all scanners offer colour and grey-scale as well as monochrome scanning and many of them can be fed with a mixture and detect which is appropriate.
- Outside of this type are models for scanning large-scale drawings, usually up to A0 where the ability to accurately feed through big areas is more important than speed.
- Machines exist for scanning bound books, microfilm, cheques and other specialist applications, but they are almost totally confined to use by service bureaux.
- It is worthwhile getting hands-on experience before selecting a scanner, and most distributors will either have a showroom which you can visit, or supply a demo machine for a free trial period.

Conclusion: Make sure you know what it is that you are going to expect from your purchase and do try before you buy.

11. In-house or outsource?

- It may be a matter of company policy to do everything possible in-house, or alternatively outsource everything possible that is not core-business. In such cases, the decision is already made, but in all others, careful consideration must be given to which way to go.
- The assumption often is that it must be cheaper to scan in-house as a service bureau does the same amount of work and has to make a profit on top. There are a number of reasons why this is not necessarily so:
    - It is the bureau's core business and it is to be expected that they would operate it more efficiently than an organisation for which it is a peripheral activity.
    - A bureau will have a variety of scanners of higher throughput than is likely to be justified in-house.
    - A bureau will have experienced staff that can take on a new job without the problem of a learning curve.
    - Scanning is not the most popular job, and whereas a bureau can rotate staff around a variety of jobs, this is not possible in-house, giving rise to absenteeism, multiple toilet visits, clandestine coffee breaks etc all leading to reduced throughput.
- It is seldom economic to set up an in-house operation to clear a backlog, as all the above apply, with the added problem of using temps (and we must assume that most organisations these days do not run with large pools of surplus man-hours). As soon as one achieves a reasonable level of competence, he or she leaves, and it's back to square one.
- For an ongoing exercise, it is not so clear-cut and if the numbers are more or less even throughout the year, then there is a good argument for setting up an in-house operation, that, if properly set up and managed, will, in the longer term, be less costly than outsourcing.
- A backlog may comprise a significant quantity to be outsourced, but the ongoing volumes may not be – most bureaux have a minimum billing charge and, in any case, will need to cover costs for collection and delivery for small quantities of work. In such cases, an in-house system would always be more cost-effective.
- Another consideration is the frequency and urgency of retrieval: if high, then it is probably not acceptable to have material stacked up for scanning until an economic quantity has accumulated to be sent away for scanning. In these circumstances, an in-house solution would not only be cheaper, but offer almost instant access to archives.
- Where there is big fluctuation in ongoing volumes, a hybrid solution may be the best answer, with an in-house facility taking the normal load and peaks being outsourced. A word of warning however – ensure your document management system can handle dual-stream input.

Conclusion: Generally speaking, back-file conversion is better outsourced and ongoing work can go either way depending on volumes and retrieval patterns. There is, however, no clear-cut formula and a lot may depend on personal preferences, local labour availability, and company policy.

12. Can we destroy our originals?

- The vast majority of original documents have no intrinsic value whatsoever and once scanned, may be disposed of without second thought.
- Many are nothing but prints from electronic documents such as emails so do not constitute original documents in any sense of the word and so there is no issue here.
- Documents that are retained to comply with the requirements of HMRC, eg invoices, bank statements, payslips etc, have long been allowed to be kept as images rather than original hard copy provided:
    - The scanning process is carried out, either in-house or outsourced, in the normal course of business.
    - Once scanned, the images cannot be altered.
    - Easy retrieval can be provided for inspection by officials.
- There are a small number of documents where the law of best evidence applies – usually those bearing signatures. In the event of a dispute going to court, consider the following:
    - One side has hard copy bearing original signatures.
    - The other has a print from a scanned image of the same document.
    - There is a discrepancy between the two.
    - There is no evidence of alteration on the original.
    - The outcome of the case hangs upon the discrepant data.
    - All things being equal, the side with the original will win.
- Some regulatory bodies and specialist tribunals still insist on hard copy being retained: these are fast diminishing, but it is well to check if your organisation is so affected and what classes of documents are required.
- There have been well-publicised cases of highly sensitive documents such as NHS patient notes that have turned up in embarrassing places like council tips: you are well advised to use a reputable secure disposal contractor. Some of these have mobile shredding units that will come to you.
- If you outsource the scanning process, any professional bureau will offer secure destruction as part of its services.

Conclusion: For the overwhelming majority, scanning to a well-organised document management system obviates the need to retain the originals which can be destroyed in a secure manner. For peace of mind, it is probably better to keep contractual documents bearing original signatures.