Misfiled your electronic documents again? If your overflowing paper files have been converted to scanned digital image format there may have been a few problems along the way. A specialist in such conversions (contact them for all your document-imaging needs) points out the important considerations there in indexing. This part of his article is reprinted by permission from Information Media and Technology 24 (6), Nov. 1991.

Document conversion may be required in a variety of different situations. One of the most important is backfile conversion, where existing documents need to be converted to digital form before the user can realize the benefits of an optical disk-based document imaging system. But whether or not there is a back file to convert, new documents will have to be converted on an on-going, day-to-day basis after the system has been installed. More specialized applications involving document conversion include CD-ROM publishing, and the capture of paper-based original information in digital form for input to computer databases or to CAD (Computer-aided design) systems.

In any document conversion project there are four key aspects to consider: preparation of the documents for scanning; indexing the documents; scanning of the documents; and quality control.

Crucially important in the document conversion environment is the issue of indexing; if you scan a document and store it on optical disk, then unless you have a good index you may never find the document again. The effect is similar to misfiles of paper documents, but with paper documents you can misfile at any time; fortunately with electronic filing, if the initial index is satisfactory, a misfile can never occur. Many DIP (document image processing) systems vendors do not highlight the importance of accurate indexing, nor provide satisfactory facilities for ensuring the accuracy of captured index data; you should check that any system you are considering will allow you to capture index data to an adequate level of accuracy. As an indication, trained data entry operators, keying numeric data only, will make about one random error in every thousand keystrokes. This means that if an index term contains ten digits, then an average of about one in every hundred documents will be wrongly indexed. We discuss below techniques for improving substantially on this accuracy level.

The main options for index data capture are: index at scan time; scan first, index later; index first, scan later.

Indexing at scan time is really only suitable for low-volume use or where the amount of index data to be entered per page is low (e.g. an average of one or two characters per page—which could of course be half a dozen characters every half dozen pages) and we would not generally recommend this for document conversion involving any sizable volumes (e.g. greater than a hundred or so pages per day). The other two options each have their pros and cons.

**Scan first, index later**

With this method, documents are scanned in batches and the digitized images held in buffer storage (typically magnetic disk). The images are displayed on high-resolution displays, which not only permit entry of the indexing data but also allow simultaneous quality control of every image. The number of workstations can be matched to the workload. The principal drawback of this method is the relatively high workstation cost, but if you are able to use the same workstations for retrieval, this may not be a major problem. Double-entry indexing can be affected by indexing at scan time then at QC (quality control) time.

**Index first, scan later (pre-indexing)**

In this method, indexing is done on batches of source documents using standard data entry techniques and cheap workstations (e.g. personal computers or conventional data entry terminals). This can be by far the most cost-effective approach if substantial volumes of index data must be captured. Double-entry indexing can be done using different operators, or second entry can be done after scanning, for example, as part of the QC process. Header sheets may be printed out and inserted into the document batches at appropriate places for subsequent control purposes. If these header sheets contain indexing information in machine readable form (OCR or bar-code), then index data may be captured automatically by the scanning system. The main drawback of the pre-indexing approach is the need for a separate quality-control process.

**Automatic indexing and full text capture**

As OCR becomes more refined, in suitable cases index data may be captured automatically from source documents. For example, if the index data is known to occur only in pre-definable areas, the scanning systems can be 'told' to apply OCR (or Intelligent Character Recognition—ICR) to the image in those areas. Optionally, the scanning systems may be able to identify which areas of the page contain text and capture full text from those areas, with the remaining areas being processed as images (mixed mode documents).
An increasing number of source documents have bar-coded index data on them, and this will be used increasingly as a very convenient aid for implementing automatic indexing.

**Methods for ensuring index accuracy**

There are a number of methods for ensuring index accuracy. Double-keying of the index data, whereby two separate operators each enter the same index data, which is then compared to see if the data agree, is a good technique; if the ‘raw’ keying error rate for numeric data is one character in a thousand, with double-entry the error rate will be one character in a million. Other methods include single-entry indexing with a visual onscreen check that index data previously captured is in accord with the displayed image; the use of index data fields which contain checksum characters; the use of bar-coding on the source documents to permit automatic capture of the index data. Particularly helpful is the use of existing computer database information to permit automatic or semi-automatic index checking; for example, if the database contains customer account number and customer name, then if the customer number is keyed, the name can be retrieved to allow visual checking that the correct number has been entered.

Ian Galbraith is Managing Director, Scanmedia Ltd, London.

---

**Indicing with death**

Under the above title, author William Gallagher wrote in the hands on word processing section of *Personal Computer World* (March 1992, pages 370-4) of his bitter experience in trying the indexing functions on various word processors. He started on a high: ‘It couldn’t be easier, right? The sales pitch for your word processor wittered on about how easy indexing was, and you confidently tell your publisher you’ll be round with it this afternoon.’ But what he found: ‘Your expensive word processor will hinder you far more than it helps—even if it helps at all...’ Word processors tend to differ in every aspect—except in being unusable.

Gallagher had used several different packages in preparing his text, and ‘to get an index I put it back through all of them’. Among his conclusions were: none of the three standard conventions for indexing (British, American, International) is followed. Adding index markers to documents is tedious, but writing a macro to do it is something most non-programmers will go to any lengths to avoid; he found ‘no clear favourite for this task’, but, ‘I know I loathe Microsoft Word’. Run-on indexes, though, seem to be available only with Microsoft Word. No word-processor he tried gave a choice of word-by-word or letter-by-letter sorting, or could make the decision to concatenate numbers or not. Alphabetization by ASCII brought problems of symbols, numbers...

Gallagher gave Microsoft Word ‘the edge on features’, as it has the run-on option, but found WordPerfect more comfortable to use, with its ‘true concordance file’ taking some of the drudgery out of the job.

Gallagher’s splendid peroration is:

Far from being the easiest and most numbingly uncreative part of writing a document, the index is a painstaking and difficult task. It’s also impossible with the software that exists now. There isn’t one piece of software that can do the job, so my only real recommendation must be, if you need an index, get an indexer (a human one) to do it.

---

**A non-intellectually produced index is no substitute...**

Reviewing the fifth edition of *Journals in Translation* (BL DSC, 1991) in *Serials* (5(1), March 1992), Craig Paterson explains:

The aim of this bibliography is the provision of a reference tool, able to facilitate the flow of information across linguistic boundaries. To meet this aim, the 5th edition applies the same tried and trusted formula of title access to three types of translation journal—the cover to cover translation, the selective translation of a journal, and consolidated articles from several journals organised on a subject basis.

The work’s internal arrangement consists of four main sequences: a principal alphabetic listing of the translated titles with bibliographic details; a keyword in context index of the translated titles; a title only list of the original titles; a listing of publishers and distributors...

complains:

The use of the KWIC indexing technique in a bibliography of this kind is far from satisfactory however. The employment of a non-intellectually produced index is no substitute for a subject with detailed subject headings and syndetic structure. Even a more conventional classified approach would have been preferable. This would have avoided the oddity of listing journals under their translated title rather than their romanised title in the head sequence, in order to accommodate the KWIC index...

and suggests:

If it is really an indexing source of translations at the article level that is required, the money would be better spent contributing to the cost of a subscription to the World Translations Index.

---

*The Indexer* Vol. 18 No. 2 October 1992