Bar codes cut indexing costs in document imaging systems

Automating document control and indexing with barcodes can dramatically improve user productivity and data integrity.

By Harvey Spencer and Daniel Borrey

Bar codes were invented in 1962 to enable automatic tracking of railroad cars. Early in 1970, they were first used on grocery cartons, and a few months later, for automobile components tracking. Widespread adoption of modern bar coding occurred because it was simpler and easier to read than optical character recognition (OCR) technology, and it guaranteed accuracy - the same features now driving the use of bar codes in document imaging.

Using bar codes in high-volume document imaging systems to automate indexes and document control can improve user productivity dramatically, while providing major benefits over other recognition technologies. Following a few simple guidelines can ensure the successful addition of bar codes to your customers' document imaging applications.

A number of creative bar-code applications are emerging in conjunction with document imaging.

In the transportation industry, bar coded tracking numbers are affixed to invoices or delivery documents, marked with notes on shortages or damage. It's critical for a distributor to receive this data promptly so a customer service department can react.

In distribution or parts ware-houses, bar codes can be affixed to a page as parts are used. The paper can be faxed and the image of the bars converted and input into an inventory control system.

Pharmaceutical companies bar code medicines and prescriptions. These can be faxed from a pharmacy to a doctor's office each time a patient fills a prescription, allowing physicians to automatically update patient records.

But beyond these implementations, the two main uses for bar codes in document imaging are batch scanning, such as on batch control sheets, and automatic indexing.

Batch scanning

Batch scanning is being used more and more in high-volume applications, defined as 5,000 or more sheets of paper to scan per day. The simplest and most common method for scanning paper is to place each single or multi-page document into a scanner's hopper, scan and view each page, and enter index fields from the keyboard. However, working this way cuts productivity, because each scan takes about 20 seconds, by the time indexes
are located and keyed, compared to the two or so seconds typically needed to scan the sheet itself. Batch scanning lets users make up the lost time.

There are two methods of batch scanning. One is used when similar documents are being imaged repeatedly, such as with Federal Express forms processing. The second is used when a group of pages represents an indexing entity, such as a medical claims form followed by supporting documents.

Batch scanning requires the creation of batch control sheets as the front of each batch. These may be used for manual control, or may be encoded so the system can recognize and control the batch automatically.

For automatic operation, two types of batch control sheets have been developed. The first and simplest is a sheet of paper preprinted with large, meaningless black areas in a predefined space. The software looks for this type of pixel pattern and automatically starts a new batch, which normally equates to a subdirectory. User's include VisionShape's AutoScan products, Sigma Imaging's Omnidesk, and TDC's DocuSystem. Normally these sheets are designed to face either end up and still work properly with the software.

A second, more complex type of batch control sheet has a meaningful bar code with a batch number encoded onto it, or in the case of litigation support systems, a case number. The software determines the start of a new batch by locating the bar and reading the batch number from the bar image.

Alternatively, a bar code can be placed on the outside of a folder. In this case the operator passes a wand over the bar before stating to scan the batch, assigning a batch or case number to attach to each image in the batch.

**Bar code formats**

Bar codes have been developed in a number of differing formats. The oldest now in use is Codabar, developed in 1972 by Pitney Bowes and stills used for Federal Express airbills, as well as in blood banks and libraries. Most document imaging systems seem to have started with Codabar. It's easy to read and relatively large; also, a clear 14-point interpretation printed above it can be verified using OCR.

Interleaved 2of5, the second-oldest bar code, is used on corrugated cartons and in the pharmaceutical industry. If possible, avoid using these codes in new applications, because their character sets are limited and can give false reads.

Code 39 was adopted by the Department of Defense in 1981 and has become the prevalent code among industrial applications and in the transportation industry. It is reliable and flexible due to a checksum and support of both alphabetic and numeric characters. Its main drawback: it uses a lot of label space, and therefore can’t be used on small documents. For this reason, it's also unsuitable for printing many codes on a single sheet.
Code 128, announced in 1981, solved many problems associated with the previous codes. It supports all ASCII characters, uses minimal integrity due to several separate message check routines. Its only drawback for document imaging is that it uses proportionally sized bars that can prove difficult to interpret, particularly when faxed.

UPC, the retail trade bar code, is rarely used in association with document imaging, nor are postal bar codes.

**Automatic indexing**

Automatic indexing using bar codes is becoming popular because OCR, the most obvious technology to use, frequently fails to deliver accurate conversions in typical forms-based document imaging applications. The forms get in the way, numbers and characters often are skewed, and documents are faint and sometimes dirty. Backgrounds and color interfere with the image, and it's difficult to locate the key fields automatically. As a result, OCR usually achieves successful recognition less than 90 percent of the time, leading to expensive index repair keying.

Bar codes solve all these issues. Software to locate the distinctive pattern of black and white lines is available. Provided that transitions and thicknesses can be identified, bar codes can be correctly converted; check digits will guarantee the conversion. With a document imaging system that can deskew, bars can appear at any angle.

A 200-dpi scan equals .005 inches per pixel. Since the thinnest bars in any bar code scheme, known as the X dimension, are usually printed from 10 mils to 0.05 mils (i.e., 0.01 of an inch to 0.05 inches), the thinnest lines normally will contain 2 pixels across. Federal Express' Codabar scanned at 300 dpi, for example, consists of wide bars and spaces of about 14 to 16 pixels wide, and thin ones of about 5 to 7 pixels wide. It's not sensible to reduce thin line width below 2 pixels because it may cause conversion errors.

However, a faxed bar code transmitted at 200-dpi horizontal resolution can be accurately converted. Paper or bar code skew is less of a factor than with OCR. If the program can effectively draw a straight line through the bar and see the differentiating black and white pixels, the decode will be accurate. Be aware of whether the bar code is vertical and the fax is in normal mode, because the scan rate is 200 dpi horizontally, but 100 dpi vertically. Due to the success in interpreting faxed bar codes, a number of applications have been developed in which bar codes are sent over the fax.

**Achieving maximum efficiency**

Although bar codes are easy to read and implement, there are guidelines for achieving peak efficiency. As we've discussed, not all bar codes are created equal. All bar codes doe have a fixed number of bard per character, so if a bar is missed, the code becomes unreadable. If adjacent bars are touching, it's possible the bar code's size will be misread; but because the table of bars is smaller than the number of possible combinations, an illegal character will be generated and the decode ejected. As a result, bar codes are significantly more secure than OCR. Substitution is virtually nonexistent and checksums will eliminate errors. Bar codes can be read in either direction and contain a different start and stop code. This can be used to recognize when a bar code is upside down, allowing a
If a standard bar code format is not in use and you need to select a bar code for document imaging, research the following before looking for a solution:

1. Is the required date numeric only or alphanumeric, with or without special characters, and will it expand in the future?

2. How much space is available for the bar code, and how many characters are required? What’s the possible position on the document? Some bar code technologies are more tolerant in areas such as reading labels placed at the edges of documents.

3. What resolution will you scan at, or what is the delivery mechanism?

4. How many misreads can you accept and does the data need check digits to ensure accuracy?

5. What's the maximum number of bar codes that you are going to have on the page?

6. What methods will be used to print bar codes, and in what form will labels be issued (i.e., preprinted, stuck on, etc.). In general, select the simplest and largest bar code an application will allow. A quiet zone or blank area of .25 inches or more before and after the bars is needed to improve accuracy. Print the translation above or below the bar code in 12-point font or above to ensure readability for
key entry. Use a check digit, create a large height-to-width ratio to allow multiple
seeks and plan on scanning at the highest resolution possible.

Note: This article can be found on the VisionShape web site at www.visionshape.com.