APPLIED INTELLIGENCE

Converting the Printed Word to Machine-Readable Text

JAMES MARTIN

This is the fourth article in a series on text management and its influence in the corporate environment.

Although managing information via computers will certainly give corporations a competitive edge throughout the 1990s, much of the data to be retrieved exists only in printed form. Before this information can be harnessed, it must be transformed into a machine-readable format.

There are two ways to convert printed matter into machine-readable form: by manually rekeying data via a word processor or by tapping new optical-scanning technology.

Optical-scanning equipment captures a printed page into a bit-mapped image, which is then converted into ASCII using optical character recognition (OCR) software. The software converts the bit-mapped page image into ASCII characters by matching the patterns of the page image against patterns stored in the software.

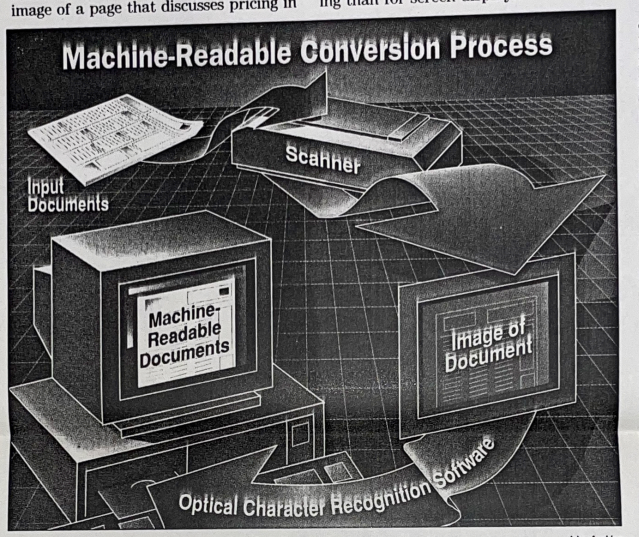
Storage of a bit-mapped image requires up to 1M bytes of memory, compared with a page of ASCII text, which requires less than 3,000 bytes. Current OCR software can recognize a variety of typefaces and font sizes, handle typeset text and flag unrecognizable characters.

Although optical-scanning systems have advanced over the last few years, OCR software is not yet 100 percent accurate, and there may be conversion errors and characters that it can't recognize. Despite these limitations, though, OCR software provides considerable benefits, primarily in the area of data-searching capabilities.

Bit-mapped page images generated by optical scanners are not searchable based on text. For example, consider the image of a page that discusses pricing in a purchasing-system reference manual. To allow users to search for any text that refers to pricing, a key word would have to be attached to the page image.

If the image was processed by OCR software, on the other hand, the resulting file could be accessed by word searches. This would let the user find the page by using a search request for any word on it.

Another key requirement of optical-scanning systems is the ability to manage the structure of text. For example, a user might wish to use different fonts for different kinds of textual material or format a document differently for printing than for screen display.



John Avakian

A technology known as document-structure markup allows for this kind of flexibility. Markup is a scheme of tags that are interspersed throughout the document file. The tags convey information about the document's structure and appearance. Markup can indicate horizontal and vertical spacing, page breaks, lists, type fonts and point size. Tags could also be used to mark repeating sections within text for easy identification.

In one scenario, a markup scheme for software reference manuals could indicate hardware implementations and version numbers, user interface sections and technical sections.

Markup can be employed for procedural purposes, such as describing how to format text on a page or what is being formatted. Using this approach, documents are not tied to a specific display medium such as the printed page.

For example, paragraphs might be marked with <PARA> at the beginning and </PARA> at the end. When the document is printed, the style guide used might indicate that <PARA> means to skip a line and include no indentation. However, if the document is being displayed on a screen, a different style guide could be used to indicate that <PARA> means to skip no lines and indent five spaces.

The best-known markup language is Standard Generalized Markup Language. SGML tags are independent of any specific word-processing package, allowing for easy transfers between packages and the text collection.

Next week I will discuss how text is indexed and queried. ■

The concepts in this article are described in a new volume, Text Management, of The James Martin Report Series. For more information on this volume, call (800) 242-1240 or (617) 639-1958. For information on seminars, contact Technology Transfer Institute, 741 10th St., Santa Monica, Calif. 90402, (213) 394-8305 (in the United States and Canada). In Europe, contact Savant, 2 New St., Carnforth, Lancs., LA5 9BX United Kingdom, (0524) 734 505.

Extenders \ *Market Thriving*

Continued from Page 68

Systems Inc. projects its revenue growth for the year to double.

Once a little-known technology, DOS extenders have come to infiltrate some of the most popular PC software applications. Lotus Development Corp. tapped Rational Systems' 286 DOS extender for 1-2-3 3.0, and both Ashton-Tate and DataEase International Inc. turned to the Rational product for versions of their databases.

Even though support from commercial software developers is strong, many customers are eying Windows or OS/2 Presentation Manager for more long-term solutions to the DOS memory crunch.

Still, many of the corporate and vertical-market developers that make up 80 percent of Phar Lap's customer base are not in this camp, said Richard Smith, president of Phar Lap in Cambridge, Mass.

The majority of these vertical-market software developers do not have the time or experience to learn graphical-user-interface programming, Smith said.

"They know about finite element analysis or rendering, but often they're not systems-type or GUI-type programmers," Smith said.

What's more, users of these specialized applications often have no need for Windows or OS/2 because these extra layers tend to diminish performance, and application-switching capabilities are useless on a dedicated computer, users said.

Take, for example, Wasatch Computer Technology Inc., a developer of high-end graphics software, which has no plans to migrate its package (which includes the Phar Lap 386 DOS extender) over to Windows. Neither Windows nor OS/2 offers the 32-bit support that Wasatch requires to generate fast graphics, company officials said.

"We need 32-bit code so we can manage 16M-byte pieces of data with reasonable speed," said Mike Ware, president of Wasatch Computer in Salt Lake City. "Windows doesn't have what we want." ■

I/F Builder \ *New DDE Support*

Continued from Page 68

processors and graphics packages, Gardner claimed. This integration would come in handy, for example, in deriving the names and addresses for a direct-mail letter drafted and printed in a Windows word processor from a mainframe customer database, he explained.

The DDE support could also be harnessed to quickly generate and update a Windows spreadsheet with mainframe numerical data to create graphs and charts, he added. "All of these functions can be integrated under the same interface so they appear to the user as one application," Gardner said.

One company turning to I/F Builder for these purposes is Information Sciences Inc. (InSci) in Montvale, N.J. InSci has used I/F Builder to create a PC-based Windows 3.0 front end to its mainframe Human Resource Management System, according to Laura Hills, InSci's vice president of product management. Called InSciVision, the Windows interface vastly reduced the number of function keys, codes and commands users need to know to navigate through the mainframe application.

"Instead of having to memorize a series of transaction codes, what users can do is click on an icon of a file folder that is labeled with the function they wish to perform," Hills said.

The interface, she said, speeds training, makes the system easier to use, and reduces keystroke errors and frustration. "This allows users to rely more on their knowledge of human resources than their ability to remember codes and transactions," she said.

Due by the end of this month, I/F Builder 2.1 will be priced at $17,500. A run-time version, called I/F Manager, is sold separately for $395 per workstation.

The software can be used to create Windows 3.0 interfaces for a variety of host mainframe systems, including the IBM 3080, 3090, 4300 and 9370 running the MVS, VM/SP or DOS/VSE operating systems.

Viewpoint Systems can be reached at (415) 578-1591. ■