GLOSSARY OF TERMS

This glossary defines terms and concepts used in this manual. For readers who are new to scanning and character recognition concepts, this glossary may be useful not only as a reference, but as a primer on optical character recognition technology, as well.

Some definitions provided here contain terms in **bold** letters. This means that these terms are also defined elsewhere in the glossary.

A

application—A software program that enables users to perform a task or set of tasks. Sometimes also refers to the use (that is, the "application") of a software program.

ASCII—American Standard Code for Information Interchange. ASCII contains codes for 128 control characters, alphanumerics, and symbols. A number of so-called extended ASCII sets exist that generally allocate another 128 codes for accented characters and additional symbols not included in the first 128.

auto page orientation—In TextBridge, a capability to correct for the rotation of the page image before recognition begins. For example, if the user were scanning a document with mixed pages (for example, most pages portrait, some landscape pages with large tables), TextBridge could perform auto-orientation on each page before beginning recognition.

auto page segmentation—In TextBridge, a capability to discern the layout of the page image, and to recognize and output text in the correct order. For example, in a newsletter, in which columns are often of uneven depths and widths, TextBridge recognizes the layout of the page and outputs text in the correct sequence. In TextBridge Preferences, you can specify auto page segmentation on or off.

B

base name—The portion of the **document name** used to identify related page image (TIFF) file names created when you use the Save Page Images capability of TextBridge. When you type a document name in the Save Page Images As dialog, the first four digits of the name are used as the base name for the page image files.

brightness—See **scanner brightness**.

C

CCITT—Acronym for Consultative Committee of International Telephone and Telegraph, and international standards organization which has created, among other things, compression standards for digital data. TIFF files stored in CCITT Group 3 and Group 4 compression standards can be recognized by TextBridge.

conversion—A software module that takes text in one format (the **input format**) and processes it to another format (the **output format**). In TextBridge, recognized text in its internal format can be converted to WordPerfect, for example (or any of a number of other supported formats).

D

DEVICE statement—In the `config.sys` file, a line that identifies, for example, a scanner's device driver to applications that may need to run the scanner.

dialog box—In Microsoft Windows, a category of user interface screen that requests interactivity ("dialog") with a user of the application. TextBridge displays a main dialog from which you can define and initiate OCR jobs.

document name—The file name you enter in either the Save Page Images As or Save As dialogs in TextBridge. The document name is automatically appended with a three-letter extension that indicates the format in which the recognized file is saved.

drop-out color—A color, or range of colors, that a scanner has problems detecting on the page it is scanning. This is typically a product of the scanner's own light source. A yellow light source in the scanner, for example, will have problems detecting colors in the range of yellow to light green.

Dynamic Data Exchange (DDE)—In the Microsoft Windows environment, a standard for sharing data among Windows applications. For example, with the proper macro in place, a word processor such as Microsoft Word for Windows can direct TextBridge to scan and recognize text, and import the text to an open Word document, from within its own menu system.

E

edit box—A Windows interface convention shown as a rectangular field in a dialog or other area of the interface into which a user can type text or in which existing text can be edited. In Windows, an edit box supports standard ways in which the user can edit information in the box. For example, in a typical Save As dialog in a Windows application, the area in which the output file name is typed is a standard edit box.

Enhanced mode—The most advanced of the three modes in which Microsoft Windows will run. The other two modes are Real and Standard. TextBridge runs only in Enhanced mode.

Expanded memory driver—A program that makes part of extended memory appear as additional expanded memory so that programs that require more than the 360K of expanded memory typically available on a DOS machine can run.

F

fax image—The representation of a page in the form of binary data (usually 200x100 or 200x200 dpi resolution) transmitted by a facsimile (fax) machine or fax modem card. Computers with fax modem cards can receive a fax image and store it on-line as a TIFF file. TextBridge can open and recognize the text from an on-line fax image stored in TIFF format.

fax modem—An external device or printed circuit board that plugs into a PC enabling the receipt and transmission of digital image data across a telecommunications (phone) line. With a fax modem connecting your PC to a phone line, you can receive and transmit document images to and from your PC.

G

galley format—The single-column format in which TextBridge outputs text recognized from multiple-column documents.

H

halftone—An image composed of differently-sized black dots spaced in such a way as to simulate the different gray tones of an original photograph or color drawing. In the Preferences dialog, you can specify that TextBridge is to ignore halftones when performing optical character recognition.

handles—Solid square objects typically in the four corners of a rectangle in a drawing package or other application which enable resizing of the rectangle. In TextBridge, you can draw a rectangular **zone** on a previewed page image to define the area of the page to be recognized. Handles on the zone enable you to resize the zone.

hypertext—A capability to traverse a textual data base (an on-line Help system, for example) in a number of different ways: by selecting a subject in an index; by stepping sequentially forward and backward; by keyword search; by context (that is, clicking on a word in text to get its definition or another screen of information about it). TextBridge uses the Microsoft Windows Help engine to navigate its built-in hypertext-based Help system.

I

input source—The origin of page images being recognized: either the scanner or a TIFF file.

ISIS—Acronym for **I**mage and **S**canner **I**nterface **S**tandard developed by Pixel Translations, Inc. ISIS is an applications programmers interface (API) for the design and development of scanner drivers. Pixel Translations and other scanner vendors use ISIS to develop scanner drivers. TextBridge supports most of the ISIS-compatible scanners available on the market today. See also **TWAIN**.

L

language pack—A component of TextBridge that enables the application to perform OCR on a document composed in a particular language. In the TextBridge Preferences dialog, you can specify that the document to be recognized is in one of English, French, Italian, German, or Spanish. TextBridge loads the appropriate language pack before beginning recognition. See also **recognition language**.

N

native user interface—In the **TWAIN** specification, the set of screen displays and keyboard controls that a TWAIN source driver provides to programs supporting the TWAIN device. For example, TextBridge runs with scanners that have fully TWAIN-compliant source drivers. However, the controls for that scanner are provided in the native UI, not in TextBridge.

noise—An errant mark on a page that can be recognized as one or more characters during OCR.

O

optical character recognition (OCR)—A technology in which binary images of character shapes are analyzed and identified as particular characters and output to a text data stream, either in computer memory or to a computer file.

OCR printer—A TextBridge application designed to work like a printer from virtually any Windows-based fax or imaging program. In the host program, you display a fax or other image containing text and use the Print command to send it to the OCR printer. The OCR printer performs TextBridge OCR, then displays the Save As dialog to allow you to save the text to a file in the desired text format.

OS/2—A graphical operating system designed by International Business Machines (IBM®) Corporation to run on Intel-based personal computers. OS/2 is a true multi-processing operating system which can run native programs, as well as programs designed for Microsoft Windows and DOS.

output text format—See **text format**.

P

page image—A binary (black and white) picture of a page stored in computer memory or on disk. Page images are scanned or read from a TIFF file and sent to TextBridge for optical character recognition (OCR).

permanent virtual memory—See **virtual memory**.

pixel—Short for “picture element,” one of many dots that make up a digital image.

preferences—In TextBridge, the settings that you can specify to control the OCR process.

Preview—In TextBridge, a capability that enables you to view, zoom, and zone a page before processing.

Q

questionable word—A word that fall below a confidence threshold built into TextBridge. During OCR, TextBridge assigns a confidence value to each word. If the value falls below the confidence threshold, and you are using the **Verifier**, TextBridge displays the word as questionable.

R

RAM disk—A part of your computer's extended memory set up to behave like a hard disk for temporary file storage.

recognition—The TextBridge process during which a page image (scanned or on-line TIFF) is analyzed, and characters and words are identified and saved as a text data stream in memory or in an on-line temporary file. Note that the TextBridge recognition engine not only performs recognition (OCR), but also performs segmentation, orientation (rotation), format analysis, and retention of text styles (bold, italic).

recognition language—The primary language (for example, English, French) in which a document is composed. In TextBridge, you can specify that the document is in one of a number of different languages. TextBridge loads the appropriate **language pack** before beginning OCR.

region—A logical block of type on a page image. TextBridge, through its **auto-segmentation** capability locates regions of text on a multi-column document, and outputs them in the correct order.

resolution—The degree of detail, measured in **dots per inch** (dpi), with which a scanner or fax machine can input an image. TextBridge can perform OCR (optical character recognition) on page images in any of the following resolutions (dots per inch): 400x400, 400x200, 300x300, 200x200, 200x100, and 100x200.

RTF—Rich Text Format, a text format developed by Microsoft Corporation, with embedded codes to describes fonts, formatting, and so on.

S

scanner brightness—A setting to determine the intensity of light the scanner projects on the page being scanned in order to lighten or darken the resulting image. Often by adjusting brightness, you can manipulate the accuracy of recognition (for example, brightening a page whose characters are tightly spaced can improve recognition). In TextBridge, you can specify scanner brightness in the Scanner Settings dialog.

scanner driver—A program that is written as an interface between a software application and a scanner. The scanner driver sends requests from the application to the scanner in a language the scanner can understand.

Scanner Setup—Part of the TextBridge OCR program group in Windows, this program is designed to enable you to load the correct high-level driver so TextBridge to run with your scanner. See also **ISIS** and **TWAIN**.

T

text format—The word processor, spreadsheet, or other file format to which recognized text can be converted and output. TextBridge supports output of recognized text to these formats:

- Ami Pro (2.0, 3.0)
- ASCII (Standard, Smart, Stripped)
- dBase IV
- DCA/RFT
- DisplayWrite 5
- Excel (Mac, 3.0, 4.0)
- FrameMaker
- Interleaf
- Lotus 1-2-3
- Multimate Advantage
- PostScript
- Prof Write (2.0, 2.2)
- RTF (Microsoft's Rich Text Format)
- Samna Word IV
- Windows Write
- Word for Windows 2.0
- WordPerfect (4.2, 5.1)
- WordStar

TIFF image—A binary representation of a page or graphic stored in Tag Image File Format, an industry-standard image file format. TextBridge can recognize text from pages images stored in several variations of TIFF, as follows:

TIFF Uncompressed (Intel header)
TIFF CCITT-3 (Intel header)
TIFF CCITT-4 (Intel header)

TIFF Uncompressed (Motorola header)
TIFF CCITT-3 (Motorola header)
TIFF CCITT-4 (Motorola header)

TIFF (Intel FAXability™ header)

When you select **Save Page Image** in the Main dialog, TextBridge automatically saves scanned page images to files in TIFF CCITT-3 Intel.

TSR program—A program designed to automatically load into memory when you start your system, or to stay in memory even after you exit it.

TWAIN—An image and scanner interface standard, complete with an API, for the development of interfaces to imaging devices (scanners, fax machines, and so on). TextBridge supports any fully TWAIN-compliant scanner or other device that connects to a PC and produces binary (black-and-white) images in a supported size and resolution.

V

Verifier—A capability that enables you to view and, if necessary, correct TextBridge recognition decisions word by word. A Verifier window similar to the Preview window shows you the recognized word and its associated image on the scanned page. The recognized word is highlighted in an edit box, allowing you to type corrections if necessary.

virtual memory—Space on your hard disk set up to simulate random access memory (RAM) on your PC. TextBridge, especially on systems with only 4Mb of RAM, must be configured with **permanent** virtual memory, a contiguous group of storage blocks on your hard disk.

W

Windows—A graphical user interface (GUI) and a host of related modules developed by Microsoft Corporation for use on DOS-driven personal computers. TextBridge runs with Windows, version 3.1 and later.

word verifier—See **Verifier**.

working directory—In Windows, when an application is installed, or at any time thereafter, you can designate a directory anywhere in your DOS file system as the working directory. You do this in Program Manager by selecting the Properties command from the File menu. For TextBridge, the working directory is a `BIN` subdirectory in the installation directory you specify at installation time. The TextBridge installation program chooses by default the working directory, `C:\TXBRIDGE\BIN`, although you can select any directory in any partition on your DOS file system.

Z

zone—In the TextBridge Preview window, a rectangular border that you can draw around a portion of the displayed page image to define the area of the page to be processed.

zoom—In the TextBridge Preview window, the capability to magnify ("zoom in") a page image to full resolution and back ("zoom out") to a resolution that enables the entire page image to be viewed.