Publishing User Documentation: PDF or HTML?

Michael Davis, Bassett Consulting Services, North Haven, Connecticut

ABSTRACT

At SUGI 23, the author presented the case for replacing paper documentation with electronic formats, and explained how to go about creating user documentation using Hypertext Markup Language (HTML). Since then, he has come to recognize that in many situations, Adobe Acrobat (PDF) files offer some attractive benefits over HTML files when publishing user instructions for SAS® applications.

This presentation will contrast the benefits and drawbacks of these two document formats. The focus will be the appearance, portability, ease of creation, and the software needed to read and print each format. The remainder of the presentation will be a tutorial on how documentation may be converted to PDF files by using either Adobe Acrobat or word processing program "plug-ins". Last, advice on some of the common traps in creating PDF files and how they may be avoided will be offered.

Please note that for illustrative clarity, all examples of Acrobat document creation are shown as they would appear on an Microsoft Windows system.

INTRODUCTION

For the purposes of preparing user documentation, HTML presents the following problems:

- documents can be altered by the reader
- appearance varies among computers
- pages are not formatted for printing
- only GIF, JPEG graphics are supported
- navigational tools not built into format
- hard for the visually impaired to read
- final layout must often be adjusted by hand

When documentation is distributed as a word processing file or as an HTML document using a file or web server, the original documents and their embedded graphics can be protected by setting file attributes to read-only. However, an author has no way to prevent any reader from copying the document to another disk volume and altering the contents. This creates the potential for mischief.

Because of their design, web browsers are customized to display HTML documents according to a series of formatting rules. The result is that the document's appearance is largely governed by the capabilities of the windowing system and the display mode being used. Issues such as resolution, window size, and available colors impact how lines of text break and how the document appears. There is no guarantee that a document will look as the author intended.

Another limitation of HTML is that the combination of the language and browser is a system intended for online viewing. The concept of predefined page breaks is foreign to HTML. Many of the formatting features built into word processors, such as headers, footers, footnotes, and page numbering are not part of the HTML standard. While every browser provides the ability to print the web page being viewed, the author has no control where long pages will break.

The HTML standard guarantees that there will be built-in (native) support for graphics in the GIF and JPEG formats. While GIF and JPEG files handle most situations adequately, what happens if your application employs a BMP, PCX, TIF, WMF, or any of the more than two dozen popular graphic formats? How does an author import a clipboard image? While software exists to convert other file formats to GIF and JPEG files, an author might not have such software available.

A table of contents and index can help readers find the relevant portion of a document quickly, regardless of whether the document is being read online or from a paper copy. While such features can be added to HTML documents, the PDF document format provides built-in support for these features, making it easier for an author to add them. It also insures that readers can access these features in a consistent manner across documents.

Those gifted with good eyesight take the ability to read a document without assistance for granted. However, the many readers who suffer from visual impairment struggle to read documents online. Some browsers allow the user to switch to larger fonts. However, unlike the typical browser, the Acrobat viewer provides the ability to magnify (zoom) any document up to 8 times its original size.

The ability of many current applications, including SAS, to generate HTML can make it easier to create documents for viewing with a browser. However, nearly every formatting tool makes certain assumptions about how the author wishes the document to appear. If those assumptions are at

odds with the author's wishes, then some manual intervention, often at the source code level, is required. Faced with the task of "tweaking" documents, many authors yearn for a tool that renders a document in paper format into its clone when converted to an electronic distribution format.

So for these and other reasons, PDF files can be a better alternative in which to render documents for online distribution and viewing.

Popularity of the Acrobat Format

In the paper that the author presented at SUGI 23 on publishing user documentation electronically, one of the advantages cited for using HTML over PDF (Portable Document Format) files was that the Acrobat Reader (viewer) was not as widely distributed as were web browsers, which were often supplied by computer vendors with new systems.

While that assumption may still be accurate, the Acrobat Reader is one of the most widely distributed pieces of software. Adobe Systems is the second largest software company in the world. Countless copies of the Acrobat Reader have been distributed via the World-Wide Web and CD-ROMs used to convey other software. Recent attendees to SUGI and regional SAS conferences have received copies of the Acrobat Reader along with the conference proceedings.

Also according to Adobe, over several hundred thousand web sites make documents available in PDF files. Among those sites is the SAS Institute web site and those of local SAS user groups. The PDF file is a de facto standard for document exchange in many industries. For example, the FDA now receives drug applications and study data in PDF files from pharmaceutical firms.

As with some of the most popular web browsers, the Acrobat Reader may be distributed royalty-free. Alternatively, recipients of Acrobat documents may download the reader for free from the Adobe web site, whose URL cited at the end of this paper.

As with web browsers, versions of the Adobe Acrobat Reader are available for use on computers running Windows, UNIX, and Macintosh operating systems. There is even a version that can be used under LINUX.

Acrobat is designed to work directly with popular web browsers as a "plug-in". When Acrobat has been installed on the user's computer and the browser is pointed to a PDF file, the document can be automatically viewed from the browser. The user might have to tell the browser through a dialog box to open PDF files with Acrobat. However, once this has been done, opening and reading PDF files from the web browser occurs automatically.

So cost and availability of the Acrobat Reader should not discourage any author from using the PDF format to distribute documentation.

The PDF File Format

The PDF file format is an open specification developed by Adobe. In fact, the details for the PDF file format are available from Adobe at a URL furnished at the end of this paper. PDF files are constructed in three layers.

The first layer consists of the text and images. It uses the imaging model of Adobe Postscript, which goes a long way to explaining why the generation of PDF files often starts with creating a Postscript file.

The PDF format employs compression algorithms, such as the JPEG, LZW, and ZIP methods. Further, it can eliminate redundant graphics for further size reduction. Fonts embedded in the document can also be compressed. By some estimates, a document that has been rendered in HTML can be reduced to as little as 20 percent of its former size when stored as a PDF.

The PDF file is optimized for good online performance. Because of progressive rendering, when a PDF document is viewed from within a web browser, it can be read before the entire document is downloaded. The text is rendered first, followed by images and embedded fonts.

The second layer of PDF files consists of navigational enhancements, such as bookmarks, hypertext links, indexes, and thumbnail views. The third layer contains basic information about the PDF file, such as the names of the fonts used and cross-referencing information for navigating within the file.

PDF or HTML?

Given this additional information, the question may arise, "Which format should I select to present my documentation? PDF or HTML? If the document is to be read online and only a casual printout may be required, then the HTML format should be satisfactory and will probably be readable by the largest potential audience.

The HTML format enjoys the advantage that web browsers automatically generate line breaks so that text fits within the available window size. Thus to read an HTML document, the user need only scroll up and down instead of having to either scroll side to side or to resize the document with the Zoom command.

The other significant advantage of the HTML format is that no additional tools need be purchased to generate an HTML document. As noted in the author's previous paper, which is cited in the Bibliography, free tools are available via the World Wide Web and elsewhere to author HTML documents. Many word processing and presentation applications, as well as the SAS System allow one to save documents as HTML.

By contrast, it will probably be necessary to purchase a copy of Adobe Acrobat, which includes the tools necessary to create PDF files if one wishes to publish Acrobat documents.

However, once the necessary software has been acquired, it usually becomes easier to generate PDF documents than HTML files. Authors can prepare documentation in their favorite word processors and convert the final drafts to PDF files with no more effort than the task of printing a paper copy.

It is important to keep in mind that the decision to use either HTML and PDF format files is not a mutual exclusive choice. Each format does some things better than the other. Neither format can completely replicate features of the other.

Best of all, PDF files can link to HTML pages and HTML pages can link to PDF files. Thus, HTML and PDF documents can work hand in hand with each other.

Getting Around a PDF

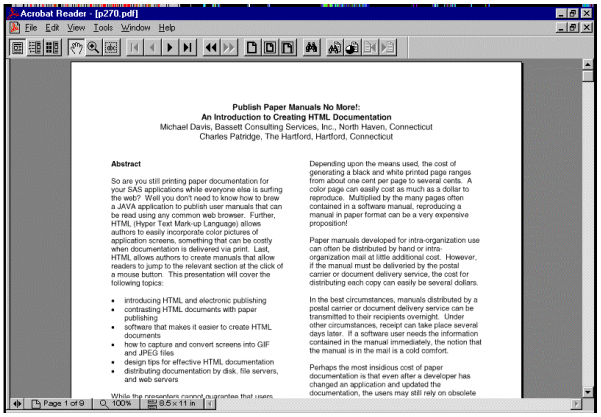

Figure 1

While use of the Acrobat Reader is relatively intuitive, some users may not be aware of all its features. When Acrobat is started and a document has been opened, one is presented with a display similar to that shown in Figure 1. *Please note that these screen shots are of Acrobat Reader 4.0 and the appearance of the icons has been slightly modified in Acrobat Reader 5.0.*

Note the row of icons immediately below the pull-down menus. The first bank of three icons, shown in Figure 2, controls the view mode. Page-only mode has been selected.

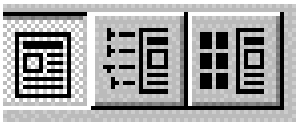

Figure 2

Bookmark View

Were we to open a document with bookmarks (e.g., the READER.PDF supplied with Acrobat) and click on the next icon, both the bookmarks and the page are displayed, as shown in Figure 3.

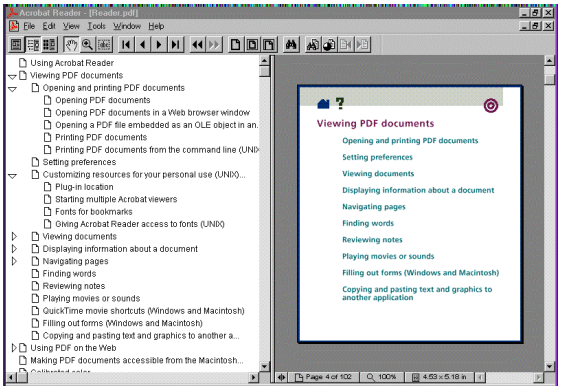

Figure 3

The bookmarks, shown in the overview pane on the left, function as a table of contents or a document outline. Click on a right arrowhead and subtopics appear. Click on a down arrowhead and and the subtopics are collapsed back into the higher level topic. Jump to a topic by clicking on it and the page appears in the right pane. Please keep in mind that bookmarks have to be added by the author for this feature to be available, using Adobe Acrobat or Distiller.

Thumbnail View

Were we to click on the right-most icon shown in Figure 2, we would see a view similar to that shown in Figure 3 except the bookmarks would be replaced by miniature versions of each page (thumbnails). Using the current version Acrobat, an author has to create thumbnails using Acrobat. Thumbnails are not as useful for documentation purposes as they might be when PDFs are used to distribute copies of SAS output.

Cursor Tools

Figure 4 shows the icons that switch which tool is controlled by the cursor.

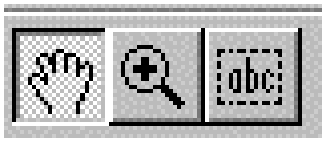

Figure 4

The left icon switches the cursor to the "hand" tool. The hand tool allows the user to "drag" the current page up or down by holding down the left mouse button while moving the cursor. When the user moves the hand over a link, the hand icon changes to pointer to the link. Clicking will select the link to which the hand is pointing.

The picture shown on the middle icon depicted in Figure 4 is that of a magnifying glass. To zoom in on a particular section of the document being viewed, click on the middle icon and then move the cursor to the portion of the document to be magnified. Click to magnify.

The right icon switches the cursor to the text tool. The text tools allows the user to copy a portion of the document to the clipboard. This feature does not work well on two column documents such as this paper. If the user selects more than one line, text from both columns is selected and copied. Also, blank spaces in computer programs are sometimes lost when copying with this tool.

Scrolling Icons

The next six icons scroll the opened document. Please refer to Figure 5.

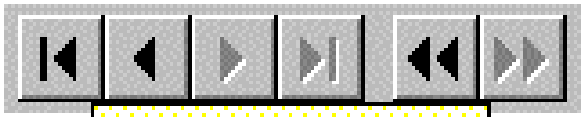

Figure 5

The icons showing arrowheads pointing to the vertical bars move the document to the beginning or to the end. The single arrowheads move the document one page at a time in the selected direction. The double arrowheads allow the user to quickly return to the previous view and then back.

Zoom Icons

Figure 6 shows the three zoom icons.

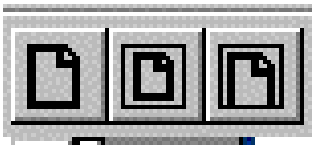

Figure 6

The first icon zooms the document to 100 percent of its actual size. The second icon resizes the current page to fit within the Acrobat window. The last icon resizes the current page so that its width is that of the window.

Find and Search Icons

Figure 7 depicts the various find and search icons.

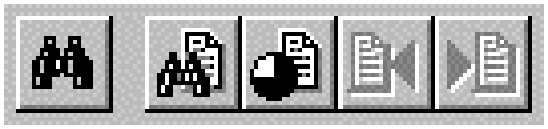

Figure 7

The binocular brings up the Find dialog, which performs a full-text search by reading every word on every page of the document. This can be slow if the document is long.

By contrast, the next icon, a binocular in front of a page, brings up the Search dialog. To search a document, the document's author must provide indexes. If indexes are available, they have to be attached to the search dialog by clicking on the Indexes button in the Search dialog box. Search is much faster than the Find command.

The Search dialog also has more features than the Find dialog. The user can search using word stems,

sounds like, and proximity. Wildcards (*, ?) and Boolean logic are permitted.

The icon showing a pie chart in front of a page redisplays the last search results list. The last two icons move the user to the next or previous item on the search results list. From the File pull-down menu, the user can further customize how the Search dialog works and returns results.

Creating PDF Documents Fast and Easy

At this point the reader might exclaim, "Enough already. How do you create a PDF document?" There are two primary ways: Distiller or PDFWriter. One method is to create a PostScript file and run Distiller.

A common question whenever PostScript is mentioned is "How do I create a PostScript file?" The author's time-proven technique is to print the document to a file using an Apple LaserWriter print driver. After all, PostScript is only a page description language used by printers and the LaserWriter is a PostScript printer.

LaserWriters have been around for over a decade and every version of MS Windows has at least one LaserWriter driver. However, some advise that a generic PostScript print driver be used to create files for Acrobat. If one wishes to follow this advice and cannot find a generic driver elsewhere, Adobe has made their own generic PostScript drivers freely available. They can be found at the URL given at the end of this paper.

To create the Postscript file, select the Print dialog box from within your favorite word processing program or other application. Select the PostScript printer, click on the "Print to File" checkbox, and click on the OK button. On the next dialog box, supply a file name with a ".ps" extension and click on OK again. You now have a PostScript file.

Creating the PDF file is equally simple. Start up Acrobat Distiller. Click on the Open selection under the File pull-down menu and select the PostScript file created in the previous step. When the next dialog box requests a PDF file name, select or enter the appropriate name and click on Save. A few seconds later, your PDF file is ready for viewing.

Please note that the Acrobat Distiller has other modes of operation, which will be covered in a later section of this paper. As of this writing, Acrobat Distiller is available for Windows, Macintosh, and some of the popular variations of UNIX.

Incidentally, if you do not have a PostScript printer and want to inspect the quality and appearance of a generated PostScript file, consider converting it into a PDF. The PDF can then be inspected or printed by any computer with the Acrobat Reader. The Acrobat Reader will render the document with relative fidelity to its appearance were it to be printed with a PostScript device.

Printing Using PDFWriter

Another way to generate a PDF file from your application is to use PDFWriter. PDFWriter is a printer driver. It is not any different than a Windows driver that might be installed for a laser or ink jet printer connected to a computer.

On Computers running Microsoft Windows, the Acrobat installation program may install macros for Microsoft Word and Excel. When the macros are installed, the user will see the selection "Create Adobe PDF …" on the File pull-down menu. Regardless, a user can always select the Acrobat PDFWriter as a printer during the Print dialog.

To create a PDF file using PDFWriter, select the PDFWriter from either the File pull-down menu or as a printer name on the Print dialog box. If you wish to change any of the settings of the PDFWriter driver, select Print instead of "Create Adobe PDF …" and click on the Properties button while viewing the Print dialog box.

Figure 8 shows some of the compression and compatibility options that you can control through the PDFWriter's option dialog boxes.

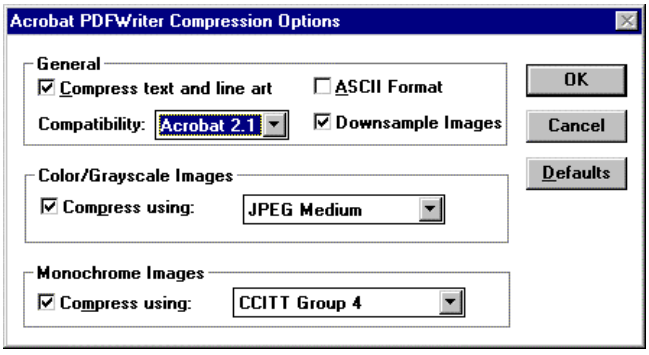

Figure 8

The default for the Compatibility is Acrobat 2.1. (3.0 in Acrobat 5.0) To use all of the compression options, Acrobat 3.0 must be selected. It is relatively safe to restrict compatibility with Acrobat 3.0 and later. Since the Version 3.0 Acrobat Reader can be freely downloaded, the recipient of a

document should be able to upgrade their version if necessary.

Another button, labeled "Fonts", brings up a dialog box showing the available fonts and list boxes to control the embedding of fonts. Embedding fonts will make the PDF file somewhat bigger and if the file is delivered over the Internet, it will download more slowly.

However, if an author uses a font that is not commonly available, the document may not be viewed successfully or print correctly on other computers. The author's advice is to use standard, commonly available fonts and test-read a document on other machines. If problems are encountered, redo the document, embedding the required fonts.

If some of the intended readers are using computers unable to handle long file names, please be sure to select a name for the PDF no longer than 8 characters.

During operation of PDFWriter, the following dialog box may appear as shown in Figure 9.

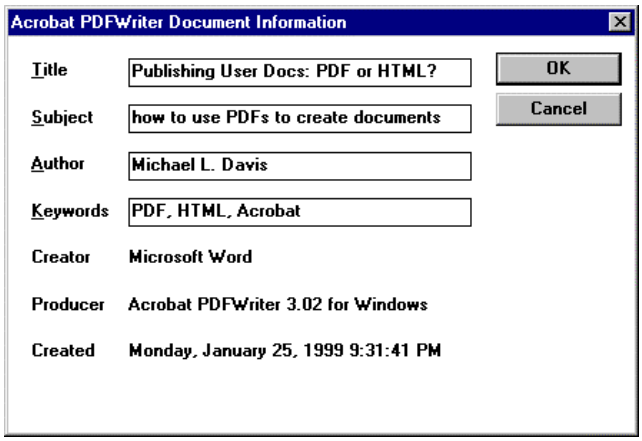

Figure 9

It is a good idea to take advantage of this dialog box as it can make identifying the content and version of a PDF document easier later on. Document information can also make the use of indexes more efficient since it facilitates searching by author, subject, or keywords.

Customizing PDF Documents With Acrobat

By using Distiller or PDFWriter, authors can convert documentation into a portable format that can be printed and viewed by anyone equipped with the Acrobat Reader. However, a long document without any navigational aids can be hard to use and is likely to be ignored. Acrobat is the complementary tool supplied to add the hypertext links, indexes, bookmarks, and other features noted earlier.

For the adventurous, Acrobat allows the author to add features that have no printed counterparts, such as movies and sounds. Similar to HTML files, PDF documents can also contain buttons and forms to support complex activities.

Optimizing Downloading From the Web

Documentation is often delivered via web servers. To speed up the downloading of PDF documents, the "Save As" or "Batch Optimize" menu selections allows the author to replace subsequent references to graphics, text, and line art with pointers to the first occurrence of those objects. Optimization also reformats the PDF so that it can be downloaded and viewed one page at a time.

To create the smallest possible PDF files, when using Acrobat to customize a PDF, use the "Save As" selection from the File pull-down menu instead of the "Save" selection. The Save selection appends the additional information to the end of the file. The Save As selection optimizes the document to minimize the PDF size. In fact, if you repeatedly use the Save selection, on the tenth save, Acrobat will prompt you to Save As.

Creating Bookmarks

Bookmarks can be an extremely helpful feature for readers attempting to find their way through even modest size documentation. Fortunately, bookmarks are relatively easy to create while viewing a document within Acrobat.

First, bring up the Bookmarks and Page view. Then select the text to be bookmarked. From the Document pull-down menu, select New Bookmark (or Ctrl-B). The new bookmark will be inserted into the Bookmark pane on the left.

Creating and Using Indexes

Supplied with Acrobat is another companion application called Adobe Catalog. Catalog allows an author to organize a collection of PDFs and build a searchable cross-reference to the PDFs, stored in a file with a PDX extension. Indexes differ from bookmarks in that bookmarks are designed solely for intra-document navigation.

To create an index, start Adobe Catalog. To create an index, select New. To update or alter an existing

index, select Open. As an example, the Acrobat help index has been opened and is shown below.

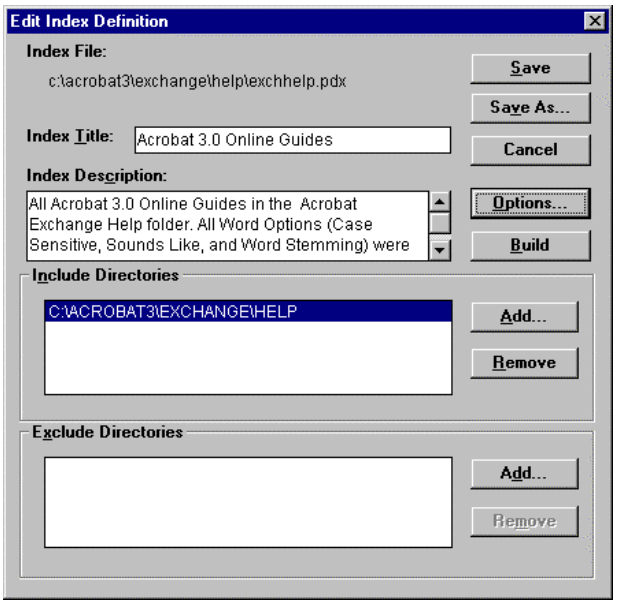

Figure 10

To create or update the index, click on the Build button. If the build is interrupted, the results of the partial build will be saved and preserved.

The author can control some of the parameters of the Build process through the dialog box displayed in Figure 11 that appears when the Options button is selected from the New or Open dialog box shown in Figure 10.

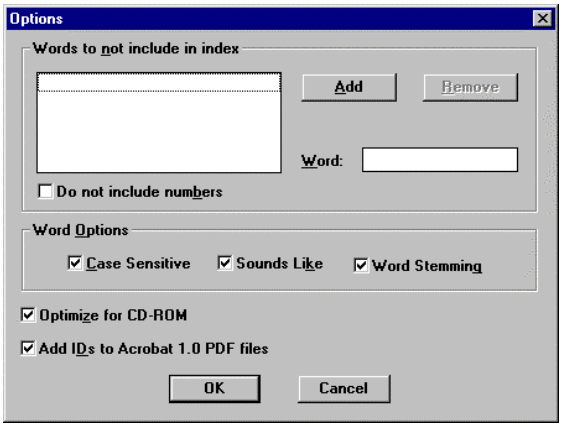

Figure 11

For the smallest indexes and fastest searches, consider editing the ACROCAT.INI file and altering the value of IndexAvailableGroupSize to 1024 files.

Securing a Document

While one might think of documentation as something to be distributed without reservation, it is not difficult to imagine circumstances where the author might wish to secure a PDF file from unauthorized copying or alteration. This can be controlled from the Security dialog box, accessed by clicking on the Security button in the upper right-hand corner of the Save As dialog box.

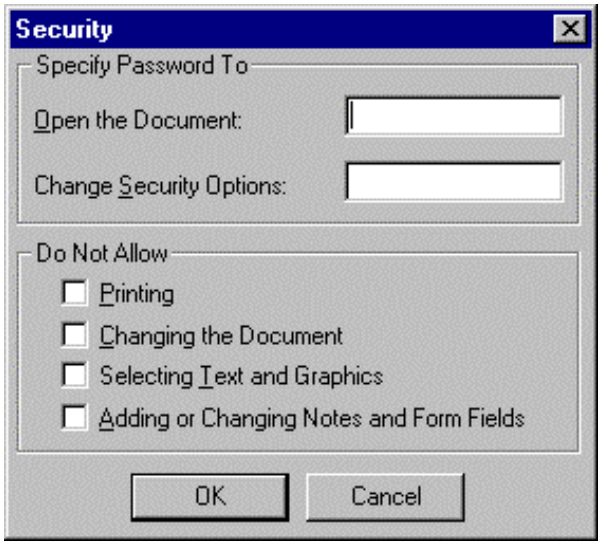

Figure 12

This dialog box allows the author to secure the document by password and to prevent unauthorized printing, copying of text and graphics, and changes.

Additional Tools

Acrobat (not the Reader!) provides some tools in addition to those described when the features of Acrobat Reader were discussed. Figure 13 shows five icons on the tool bar, which selects some of these additional tools.

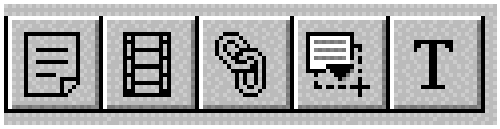

Figure 13

The first icon allows the insertion the electronic equivalent of "sticky notes". The second icon allows the insertion of movie clips, which can be viewed only, if the target computers support this feature. The middle icon allows the insertion of hypertext links. These links can point to other portions of the PDF document, other PDF documents, or an

Internet address. The fourth icon selects the form tool. The last icon selects the touch-up text tool, which allows an author to edit text context and appearance. This can be a handy feature when the author does not have the original document used to generate the PDF file.

Thumbnails

Thumbnails (illustrations) are probably not as useful to the authors of computer use instructions as they are to the distributors of other types of documents. However, authors of PDFs should know that Acrobat Acrobat can create thumbnails from the Document pull-down menu or during the Batch Optimization selection from the File pull-down menu.

Thumbnails take additional space and unless there is a need for them, they should not be created. Thumbnails do have one feature that may be helpful to the creators of documentation. Similar to MS PowerPoint, the thumbnails can be used to reorder pages in a PDF document.

Articles

The ability to organize content into articles is another feature of Adobe. Articles allow the author to specify a reading sequence for information presented in snaked column format, such as with a newspaper or telephone book listing. Select Article from the Tools menu to work with this feature.

Page Layout

Acrobat Reader and Acrobat allow the reader to scroll through a PDF document in one of three different ways. They are:

- single page at a time
- continuous layout
- continuous facing-page layout

The difference between continuous layout and continuous facing-page layout is that the facing page layout shows the pages paired instead of in a single vertical column.

Acrobat allows the author to select the default page layout for a document. However, such selections work again the design principle of Acrobat, which is to let the reader select how they can mostly easily view a document.

Web Sites URLs to Note

The following web site URLs (Universal Resource Locators) can be helpful to those who wish to increase their knowledge about PDF files. First, Adobe maintains a very comprehensive web site for Acrobat, including links to other sites. The Acrobat site is at

http://www.adobe.com/prodindex/acrobat/main.html

The PDF Software Developers Toolkit (SDK) is found at (please excuse the wrapping of the URL):

http://partners.adobe.com/asn/developer/acrosdk/docs/contents.pdf

Finally, all types of information and links to other web sites about PDFs can be found at:

http://www.pdfzone.com

Last, when you need to install a PostScript driver on a computer and are unable to unearth the install disk or CD-ROM, you can obtain a universal PostScript driver from Adobe. The URL for the page featuring the various downloadable drivers from Adobe is (again please excuse the wrapping):

http://www.adobe.com/supportservice/custsupport/LIBRARY/pdrvwin.htm

With these URLs in one hand and a copy of Acrobat in the other, one should be ready to take good advantage of the features of PDF files.

CONCLUSION

The author hopes that this paper aids his peers in selecting the most appropriate tools to prepare documentation. He also hopes that this detailed description of Acrobat and the features of the PDF file format encourages readers to use the PDFs to distribute documentation.

Please note that this paper is still being revised and improved. A link to the most current version in Adobe Acrobat format (PDF) has been posted to the World Wide Web. A link to it can be found at the URL http://www.bassettconsulting.com.

BIBLIOGRAPHY

Davis, Michael and Patridge, Charles (1998), "Publish Paper Manuals No More!: An Introduction to Creating HTML Documentation, " *Proceedings of the Twenty-Third Annual SAS Users Group International Conference,* 23, 1508-16.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the help and encouragement of his friend, Mike Zdeb, New York State Department of Health in preparing this paper and all things PDF. He also wishes to thank Adobe Corporation, which furnished a copy of Acrobat 5.0 to review in updating this paper for NESUG 2001.

CONTACT INFORMATION

The author may be contacted as follows:

Michael L. Davis
Bassett Consulting Services, Inc.
10 Pleasant Drive
North Haven CT 06473-3712
Internet: michael@bassettconsulting.com
Web: http://www.bassettconsulting.com
Telephone: (203) 562-0640
Facsimile: (203) 498-1414