

Why PDF is everywhere

© By Tony McKinley, first appeared in INFORM, the journal of AIIM, 9/97

In an interview for this article, John Warnock, Co-founder and CEO of Adobe Systems, Inc. defined the Adobe Acrobat Portable Document Format: "PDF is an extensible form of paper, a hypermedia that is device independent, platform independent, color consistent and it is the best universal transmission media for creative and intellectual assets."

Since beginning development in 1991, PDF has been widely accepted across many environments. In Document Management, PDF is the rendition format of choice for published documents, widely embraced by such major vendors as Documentum, FileNet/Saros, Open Text and PC DOCS. In Information Retrieval, full text searching of PDF content is supported by Excalibur Technologies, Fulcrum Technologies, PLS, Verity and others. And in this age of software distribution on CD-ROM, many vendors ship digital documentation on disk in PDF format. Beyond that, information from such diverse sources as the Government Printing Office, popular newspapers, financial analysts and many others is published on the Internet in PDF.

Beyond business and governmental users, PDF acceptance has spread to the everyday user on the Web. Perhaps the most dramatic example of this is the availability of tax forms on the Web, where the faithful reproduction and print features of PDF allowed downloading of forms rather than trips to the Post Office or other sources to pick them up. One measure of the demand of such online availability of forms is the fact that Adobe's Web site downloaded 98,000 copies of the free Acrobat Reader in one day on 4/15/97, allowing people to print the forms on demand on their own printers.

Adobe's Acrobat format is similar in world-beating common sense to Adobe PostScript. Just as PostScript liberated page creation packages from page output devices, which in the original environment meant hardwired printers to typesetters, so that any author or publisher could create one document that would be usable by the greatest number of clients, agents and casual workers. In the case of PostScript, the universal language was between mostly Mac's and PC's. In the bigger world of universal documents, PDF is viewable, printable and share-able between over thirteen platforms. PDF is a much more self-contained, and thereby transportable, "universal" medium than PostScript. More importantly, the Acrobat Reader is specifically adapted for most user environments.

"The history of PDF started in 1991," according to John Warnock, "when it occurred to me that everybody has been talking about a 'Paperless Office' for decades... and we envisioned it all as ... plain ASCII ... green letters ... all upper case" Warnock chuckles, as do we all who have been here since before there were PC's, before there were word processors, even. "The Paperless Office has always been the Holy Grail," Warnock continues, and "I looked around and saw computers on top of all the desks. This meant we could grab all the content as printer output, and this was critical to solve the font problem."

According to John Warnock, "The three seed inventions that led to Acrobat were the focus on the print stream because all of these desktop computers were connected to printers, the ability to capture the PostScript which offered device and platform independence, and finally we needed to make synthetic fonts so we wouldn't have to ship the fonts with the documents. By 1992,"

Warnock continues, "we had a prototype that could display on screen fast, but we realized we had to do the file structure exactly right. We used our best computer scientists to build a robust file structure." Taking this thinking one step further, Warnock says "We looked around at 20 jillion legacy documents on paper, and thought it would be great to capture them in this new format. So, in 1993 we bought two OCR companies and re-wrote recognition programs from scratch, mostly using the engineer's expertise from these acquisitions."

In Warnock's view, "The critical inventions in the Capture product was that unlike OCR, we didn't use tildes or asterisks to signal uncertain recognition, we put in a bitmap for any possible mistakes. This potentially eliminates the editing required in scanning, and the file is still readable and searchable. Of course, we are currently working to make this a jillion times better." Another result of this approach that is immediately obvious to any experienced OCR user is that the output of Acrobat Capture looks just like the input pages, including graphics, formatting, signatures and other elements that have historically been difficult if not impossible to retain.

"So, in the sequence of history," Warnock concludes, "we had OCR in 1993. This was way before the Web, and we had the perfect platform for the Web. The rest is still in the process of making history." When pressed for future Adobe directions, Warnock states "We guarantee that Acrobat will stay cross-platform, stay stable, and be a reliable medium for archiving. We are continually adding features to make it a repository for information."

PDF in Document Management and Information Retrieval

In Document Management, a major challenge is supporting many different document formats so that all users can access all information. Historically, every user needed to have the creating application installed locally to use the documents. This creates a huge burden not only in software costs and installation, but more importantly in the requirement for all the users to be familiar with a large number of applications. Early approaches to this requirement included bundling a viewer to handle all the various formats, which typically provided limited annotation and re-use capability. Another alternative on Web-based systems is to transform all documents to HTML on-the-fly. While this approach is feasible on very simple documents, it often leads to degradation of the presentation of the information, as discussed below.

With the adoption of PDF as the distribution format of choice, all users only have to learn one user interface, namely the freely distributed Acrobat Reader. In this situation, the document management system typically generates an automatic PDF rendition of a document for end users, while authors retain the original document in the creating application. The documents maintain their full presentation and print capability in PDF, and a variety of markup tools are available for highlighting, redlining, and annotations.

The structure of PDF files also offers additional benefits in Doc Info fields which provide Author, Subject, Title, Keyword and system fields such as Date Created and Date Modified. Advanced navigation is available in the form of easily generated bookmarks, thumbnail views and Web links. The Acrobat Reader provides for text searching, page turning and scrolling through files. Verity, Inc. offers a freely downloadable search engine called SearchPDF for Web Servers, which allows large PDF collections to be served over the Web and Intranets with the full range of Verity search features on both the Document Information fields and the contents of the PDF files. Other leading Information Retrieval vendors bring their special capabilities to text search of PDF, including Excalibur Technologies, PLS, Fulcrum Technologies and others.

Converting Paper - Acrobat Capture compared to OCR

The greatest difference between conventional OCR and Acrobat Capture is that the former has primarily targeted output that will be edited for final use, whereas the latter is designed to reproduce the entire presentation of the scanned page, including everything from graphics and signatures to the fonts. On multi-page documents, OCR results will often flow over the original page endings and column boundaries. Acrobat Capture's strong suit is maintaining the layout of the original page, with the added benefit described above of inserting snippets of images of questionably recognized text. In many cases, Acrobat Capture combines the benefits of image retrieval with full text searching. In addition, the PDF output is suitable for distribution on CD, Intranets and the Web, while OCR output is typically in word processing formats or a barebones HTML file which requires editing for full functionality.

Many otherwise viable OCR conversions are killed by the time and expense involved in editing the output. No matter how fast the actual OCR process, cleanup typically proceeds at human speeds of just tens of pages per hour. With the image snippets, Acrobat Capture offers the option to skip this stage. While this may be philosophically or functionality impossible for some applications, it must be noted that Capture still retains its best interpretation behind the image and the overall accuracy is comparable to OCR. This means that text searches will often find the suspect terms, while the document is completely readable and printable without editing. The merits of this argument can only be determined for each application.

Comparing Web formats - HTML and PDF

The fundamental difference when comparing HTML and PDF is the design intent of each format. HyperText Markup Language (HTML) is designed to deliver content to any Web connected computer. The 'HT' of HTML refers to the HyperText Transport Protocol which is the heart of the Web, the ability to instantly jump from one document to any other document or site on the World Wide Web. However, due to its nature of universal accessibility, HTML lacks precision in describing the presentation of pages. For example, in word processing or desktop publishing, we are all familiar with precise definitions of the fonts we are using, such as "Times New Roman, 12 point" or "Arial Bold Italic, 18 point." Because HTML is designed to display in any browser, and the user of the browser can set the fonts to their preference, font sizes are simply described as Big, Bigger, Biggest and Small, Smaller, Smallest. While Cascading Style Sheets have recently been proposed to address this inadequacy, they are not expected to be widely adopted until late 1998. At the moment, the two most popular browser vendors, Netscape and Microsoft, have developed independent, non-compatible extensions so that HTML page designers are still faced with the choice of designing primarily for one or the other. In addition, because HTML is a display-based presentation, precise control of printing devices is out of the question for complex documents.

On the other hand, harking back to Warnock's comments that PDF is designed to capture the print stream of document creation functions, Acrobat documents look and print the same across all platforms. The difference between this approach is dramatic on such common documents as spreadsheets. In PDF, the columns and tabs align as predictably as they would on your laser printer, whereas in HTML the lack of precise placement means that columns often wrap and lose their presentation and even readability. If you take this one step further, to more sophisticated documents from desktop publishing applications like PageMaker and Quark, HTML can't

approach the precision of color and sophisticated layout of PDF. In this case, PDF is literally equivalent to digital PostScript.

+++

Tony McKinley is a principal of Intelligent Imaging, and he has been dedicated to scanning, recognition and information retrieval since 1983. He is the author of the recently published book "From Paper to Web" by Adobe Press, which is available in linked, searchable PDF format at the companion site at <http://imagebiz.com>. Contact him at tonymck@imagebiz.com, or phone 610.647.5570.

(This article was published with First Serial Rights in INFORM, 9/97. Copyright and all reproduction rights are reserved by TM.)