

PDF: A Look Inside

James C. King
A Principal Scientist
Adobe Systems Incorporated
San Jose, California

Introduction

The Portable Document Format, or PDF, was first introduced by Adobe® Systems Incorporated in June 1993 with the announcement of a suite of products under the name of Adobe Acrobat®. More recently, PDF has become the focus of great attention as a key electronic format within the professional publishing industry and as the way to presented printable material on the World Wide Web.

The primary intent of this presentation is to remove some of the mystery of what is within a PDF file. A series of example PDF files will be examined showing how text, images, and graphics are represented within a PDF file. The general organization of a PDF file as a collection of objects will be exposed. Little to no prior knowledge of PDF is required to follow along.

Page Content

The format for PDF files breaks naturally into two classes: 1) the material representing “page content” or those directives that contribute directly to causing marking on a rectangular canvas (page or screen), and 2) the auxiliary and structural information that assists dynamic viewing or helps organize the content into a complete “document.”

The first set of examples focus on the content. Both PDF and the PostScript® Language share a common “imaging model” that includes text -- using outline font technology, graphics -- using lines and Bezier curves to form paths that can be either “filled” or “stroked”, and image sampled data -- used to form pictures. All of these imaging primitives are woven into an imaging-model cloth using flexible positioning primitives, sophisticated color directives, and coordinate space warping mechanisms that include translation, rotation, and scaling. A quick glimpse of each of these is provided by example PDF files. These content examples also serve to illuminate the equivalent functions of PostScript since it and PDF share the same imaging model.

Auxiliary and Structural Information

Further examples show how information beyond the page content is included within a PDF file including: bookmarks or table of contents, thumbnails, links, annotations, and forms fill-in information. PDF has at times been referred to as “object oriented PostScript” and the object structure of a PDF file is a basic lesson learned from all of the examples. The organization of PDF files into objects is used to maintain a page independence that allows the pages to be accessed and processed in any order. This is crucial to document browsing and dynamic linking, but also has been deemed crucial to the prepress operations needed to prepare material for various production output devices and include imposition and trapping.

Expressiveness and Flexibility of PDF

The Portable Document Format is based on an intriguing notation that is both powerful and flexible. As with the imaging model, the roots of the PDF syntax and low-level semantics derive from the PostScript Language, but the key mechanisms that allow PDF files to be considered as a collection of related objects

and to be accessed randomly are PDF's own. The notion of a "dictionary" where "keys" can be looked up to find their "values" provides a cornerstone of flexibility for PDF. Dictionaries in particular, but PDF in general, allows for a program to find the key elements of what is desired and to ignore those components or parts that are not of particular interest at this time. Again, introduction to these concepts by example is part of the presentation.

Availability of PDF Technical Information

In June of 1993 the Addison-Wesley Publishing Company in collaboration with Adobe Systems Incorporated published the definitive book called "Portable Document Format Reference Manual" describing in technical detail the PDF version 1.0. Since that time revisions to this technical specification have been available to the public directly from Adobe Systems Incorporated. The current version of PDF is 1.4 which was announced and documented by Adobe in November 2001 along with the availability of Acrobat 5.0. (That document can be found on Adobe's World Wide Web site within the technical documentation available for developers at <http://partners.adobe.com/asn/developer/technotes/acrobatpdf.html>.) From the start, Adobe has put no restrictions on other people or companies ability to read or write PDF files. In fact, the Adobe Solutions Network (ASN) provides extensive software support for developing PDF oriented products for a modest annual membership fee.

Availability of this Presentation and Supporting Materials

The example PDF files and the main PDF file used during the presentation can be found on the Adobe's World Wide Web site at:

<http://kiosk.adobe.com/users.jking>

A running commentary of the presentation organized slide by slide is also available there.

Biography

James C. King
A Principal Scientist
Adobe Systems Incorporated
San Jose, California

Dr. King is one of the people responsible for the vision, architecture, design, prototyping, and ultimate development of new products and new features for existing Adobe products.

Jim King joined Adobe in 1988 and until 1996, was the Director of Adobe's Advanced Technology Group (ATG). He is now a member of that group.

Prior to joining Adobe Systems, Dr. King was manager of I/O Systems Laboratory (IOSL) at the IBM Almaden Research Center where he was responsible for guiding research projects dealing with advanced printers, scanners, and displays.

Dr. King received a Ph.D. in Computer Science from Carnegie-Mellon University. He is a member of the ACM, IEEE Computer Society, the Seybold Conference Advisory Board, a board member of the IS&T (Information Sciences and Technology), and on the San Jose Public Library System Master Plan Taskforce. Dr. King is an inventor on numerous patents.