STATUS OF CGATS.12

USING PDF FOR DIGITAL DATA EXCHANGE

Stephen N. Zilles*

Keywords: Data, Electronic, Files, Standards

Abstract: The Portable Data Format (PDF) is a format for representing composited electronic documents for the purpose of exchanging the document between a sender and a receiver that may not have had any prior communication. PDF provides an object based representation of the content of the electronic document; that is, there are different representations for the objects, the text, geometric graphics and raster images, of which the document is comprised. These object based representations efficiently encode the various object types and also provide a reproduction-device-independent representation of the digital data. The Committee for Graphics Arts Technologies Standards (CGATS) is developing a standard, CGATS.12, for the exchange of digital data using PDF. This work is motivated by the desire to allow the electronic transmission of the creative work of the graphic artist throughout the production workflow to final production either as a final image on media or a surrogate for that image, such as a printing plate. To avoid dependence on the set of applications used to create a graphic presentation, a standard format is necessary to allow transmission of the digital work through prepress shops to the publisher and on to the printer. The DDAP (Digital Distribution of Advertising for Publications) Association has been instrumental in encouraging and supporting this standardization. This paper provides an overview of PDF and a PDF workflow and describes how PDF can be used in conjunction with raster based workflows.

Background

There have been two approaches taken to the preparation of digital, electronic representations of graphics arts presentations. In one approach, *raster-based*, the entire presentation is represented in terms of raster images, both those scanned from traditional sources, such as

* Manager of Standards, Adobe Systems Incorporated, 345 Park Avenue, MS/W14, San Jose, CA 95110-2704, Email: szilles@adobe.com

photographs, and synthetically created images such as text composed on a computer system. In the other approach, *object-based,* the representation of the presentation is based on the type of object, such as text, geometric graphics and raster images, being represented. For example, text is represented as character strings plus a font to be used to render the character string; geometric graphics are represented using commands to draw the graphic and raster images are represented as rasters. Because the second approach uses distinct representation approaches for the object types within the document, the total representation is quite efficient and is independent of the resolution of the (often unknown) reproduction device on which the digital document is to be realized.

A raster-based format in which all of the "objects" have been rasterized; that is converted into some portion of an array of color picture elements or pixels at a resolution suitable for the intended reproduction device. Because the raster representation is closer to the direct inputs of the reproduction device, the process of generating film, plates or final impressions is simpler and has a more predictable time scale.

Both object-based and raster-based formats complement each other: a raster-based format provides the tightly bound representation for which most of the rasterizing decisions have been made and an object-oriented format provides as representation that is more flexible, which can be adapted to more kinds of reproduction processes and resolutions. The flexibility of object-based formats also provides a greater capability for correction of errors in the content when that is necessary and it provides for personalization, such a local address lists, bingo card numbers, of the content of an advertisement.

Publications and other presentations may combine material distributed in either format. For example, a publication may have its editorial content in an object based format and have two partial page advertisements, one in an object-based format and the other in a raster-based format. This requires that both raster-based and object-based formats can be combined to produce a "final" presentation.

The raster-based approach is the subject of the companion paper by David McDowell in this volume on the status of TIFF/IT. TIFF/IT is a standard raster-based format which is based on the output formats used by Color Electronic Pre-press Systems. This paper presents an overview of PDF which is an object-based format and the basis for CGATS.12, a standard for PDF eXchange or digital pre-press data, also known as PDF-X.

Object-based Exchange formats

There are several widely used, object-based exchange formats based on what is now called the Adobe Imaging Model. This imaging model is a powerful way to represent text, graphics, and images in a coherent and consistent manner. It was originally implemented in the PostScript Language, which is an exchange format for printing and is now at the core of a wide range of printing and publishing technologies, including the PostScript Language, the Portable Document Format (PDF), and a number of application products.

The PostScript Language

The PostScript Language has become the industry standard for interchanging printing files for producing high-quality output. Central to the success of the PostScript Language is the fact that it is a completely device-independent page description language. This means that the same file can be printed on desktop laser printers that cost a few hundred dollars or on high-end printing systems that cost hundreds of thousands of dollars, and the document will always print with the highest quality achievable by the particular output device.

Another important aspect of the PostScript Language is its imaging model: the model by which the text, geometric graphics and raster data is imaged onto a medium. The imaging model has two components: a set of operations for painting colored regions onto the background and a graphics state which controls how the painting is done. The power of the model comes from the kinds of curves and areas that can be modeled and the level of control that the graphics state provides. Over its more than 10 years of use, this model has been shown to be both robust and powerful. Most electronic graphics arts presentations can be and are represented in the PostScript Language.

Also important to the success of the PostScript Language is that is has been documented in a publicly available form since its creation in 1984. The PostScript Language Specification Manual contains permission for anyone to implement to the specification. This has allowed multiple vendors to provide implementations and has encouraged competition to provide the best quality and highest performance.

The PostScript Language is both a page description language that describes the format of a printed page, and a general-purpose programming language. The programming language part of the PostScript Language contains procedures, variables, and control constructs that must be interpreted to render a page description. The programmability of the PostScript Language has both positive and negative aspects. On the positive side, it provides a flexibility that has made it possible to adapt PostScript to almost every document production environment. On a more negative note, the procedural nature of PostScript files means that they must be interpreted in the order in which they are streamed into an output device. One reason for this is that PostScript files are not inherently page independent; imaging operations performed on the second page of a document may depend on graphics state settings that were established on the first page.

The Portable Document Format

The Portable Document Format (PDF) was created to take the power of the PostScript Language beyond just printing. PDF and the software used to create, display and print PDF is also based on the Adobe Imaging Model. It was developed so that users could view and manage documents on-screen in a device- and application-independent manner. With PDF, documents can be created in virtually any application, on any platform, and easily converted to PDF, where they retain the full range of high-quality typography, graphics, images, and color. PDF files can be viewed, navigated, searched, printed, and archived in Macintosh, Windows, ® and UNIX ® environments[1].

PDF retains the Adobe Imaging Model, but removes the programming language portion of PostScript. This makes imaging PDF a simpler, more rapid process. PDF also explicitly represents the structure of the document: the content of the pages (and parts of pages) and the resources required to image these pages are all separately represented in the PDF file. With this structural information, each page can be independently imaged and the data required to do that imaging can be efficiently accessed. (There is even a form of PDF, linearized PDF, that is optimized for incremental access to the parts of the file necessary to image a particular page. Linearized PDF can be used to provide efficient display of randomly accessed individual pages of a document over a network such as the World Wide Web.) To reduce the size of PDF files, compression techniques are built into the format to compress the contents of pages and the resources.

PDF for Prepress Digital Data Exchange

From early on in the process of standardizing the exchange of digital prepress data, it has been realized that there is a market need for both raster-based and object-based exchange formats. For example, the requirements prepared the association for the Digital Distribution of Advertising for Publications (DDAP) refer to both the work on ANSI IT 8.8 using TIFF and the work using the PostScript Language. In 1993, CGATS SC 6 was created to develop standards to satisfy the DDAP requirements. The early object-based work undertaken by CGATS was based on the PostScript Language and Encapsulated PostScript (EPS) files. When PDF 1.0 was released in June of 1993, it became clear that PDF was a better solution than EPS for prepress exchange because its simpler, more rigorous structure made rasterization easier and freer from anomalies that are sometime experienced with EPS files.

PDF 1.0 did not, however, have the full functionality of the Level 2 PostScript Language; it lacked some of the features needed for Graphics

[1] using, for example, the free Adobe Acrobat software (available on http://www.adobe.com).

Arts uses. Work on an object-based standard using EPS began in 1993, as this work progressed, it became clear that it would be difficult to get agreement on how to constrain EPS suitably for reliable exchange. During this period, PDF was being extended to have the features required for graphics arts work. In 1995, it became clear that PDF would provide a more robust solution than EPS and the work of CGATS SC 6 shifted to use PDF as the reference for CGATS.12. With the publication of PDF 1.2 in November, 1996, almost all of the features required for graphics arts usage were part of the PDF specification.

Some of the additional benefits of using PDF for prepress digital data exchange are editability and improved archivability. Because of the structure and object-based nature of PDF files, it is simpler to construct an editor that can make necessary, last-minute changes to the file. The textual data is stored in PDF as strings rather than images. This facilitates editing the content of the string to correct errors. The structure of PDF also allows parts of the file to be identified as replacement areas which would be varied to allow local vendor names and address or bingo card numbers to be added for a particular use of the file. Also because the text is still in string form, the PDF files can be indexed simplifying the search for a particular archived file based on the content of that file.

PDF and Prepress Workflow

One reason for choosing PDF as the basis for a digital prepress data exchange format is the recognition that PDF was also a better basis for the desktop publishing prepress workflow. Today, most documents are delivered to prepress or print shops in the authoring application format. Once received, the file enters a workflow process based on the PostScript Language or native application formats.

Although the PostScript Language was initially developed as a language for describing pages and controlling printers, but its flexibility enabled it to become the data format for carrying prepress and production information as well. But, flexibility is a two edged sword; it can also lead to unpredictability, in part because so many different applications generate PostScript Language files in so many different ways and PostScript Language page descriptions can be arbitrarily complex. A typical imposition application may have to understand 200 or more different application versions of PostScript output. It is not uncommon for prepress application developers to spend half of their development time just keeping up with the latest application output streams.

In contrast, PDF files are highly structured . A PDF file can be thought of as a database of objects with direct access to each object, and each page of a PDF document is independent of the others. If a prepress application uses PDF files instead of the PostScript language as its input and output, it is able to directly access the information needed and incrementally update the file. The prepress application also has just one format to understand—PDF. The apparent arbitrariness of PostScript technology is eliminated, so PDF provides the foundation for a print production system that delivers consistent, predictable results. A PDF file delivers the single "digital master" for use in electronic, printed, and mixed workflows, ensuring the highest fidelity across all media.

PDF 1.2 includes the features necessary for PDF files to work seamlessly in production printing for color and monochrome workflows. The full functionality of the Level 2 PostScript Language, including the features for high-end printing, such as control over screening, separations and image replacement is representable in PDF 1.2. This allows a PostScript Language file to be converted into PDF 1.2 and, where necessary to use existing PostScript Language devices, can be converted back again to the PostScript Language without loss of functionality, without loss of visual fidelity. A benefit of delivering documents as PDF files is that a PostScript Language file that has been created from PDF tends to print more reliably than the original PostScript file. In the conversion to PDF, the arbitrariness is removed from the file, so that when it is converted back to PostScript technology it is more tightly structured. Some of the features of PDF 1.2 that make these conversions possible are:

> The graphics state set in a PDF file has been extended to include relevant device-dependent parameters. This "extended graphics state" allows the specification of stroke adjustment, overprinting, black generation, undercolor removal, transfer functions, halftone screens, and halftone phase.
>
> The information in Open Prepress Interface (OPI) version 1.3 comments can be represented in PDF files, enabling OPI image replacement to be preserved in the PDF file. Very large high-resolution images can be stored separately from the PDF file itself, allowing small files to be maintained and routed with the large images replaced at print time.
>
> The full set of Level 2 PostScript Language color spaces is in PDF 1.2. This included adding separation and pattern color spaces. A separation color space can be specified for any separations (spot colors or process colors) to be produced by a given device. If the output device does not support the specified separation, it will use an alternate color space (specified in the PDF file) for predictable behavior. A second new color space, the pattern color space, allows the printing of PostScript language patterns.

A typical scenario using PDF for document delivery would be: a graphic designer creates a design for an brochure. He obtains the images needed in photographic form and sends them the printer for scanning and storage of the high resolution images. The printer returns EPS files with

FPO (for placement only) images and OPI comments describing the high resolution files. The graphic designer creates a document in a page layout program, includes the FPO images received from the printer, and then outputs the document to a PDF file. Any printing control features specified in the authoring application are maintained. OPI comments specified in the EPS files are included in the PDF file so high-resolution images can be added back into the file before going to press. The (small) PDF file is then transmitted to the printer, reducing the time and effort it takes to transfer. When the PDF file is received at the printer, initial preflight is streamlined because all of the components are in one neat package and viewable on-screen. The document is generally output to a PostScript language file at this stage, maintaining the print controls originally specified in the authoring application. It is routed through a prepress workflow, high-resolution images that remained at the printer are replaced in the file, and the document is output to final film, plate or paper.

Commercial experience with PDF Ad Delivery

The Associated Press (AP) operates a digital advertising delivery service called AP AdSEND. This is one of several services providing for the delivery of advertisements in PDF format. The AP delivers advertisements via satellite to newspapers throughout the United States. Using PDF provides tremendous advantages for retailers and advertising agencies because of the cost savings, ability to make important last-minute changes, faster time to market, and higher reproduction quality.

At present, AP AdSEND supports more than 1,300 newspapers receiving advertisements in PDF format from more than 400 major advertisers. The volume is now more than 70,000 full-page advertisements per month; and up to 4,600 advertisements per day at peak times. Up to this year, almost all of these PDF format advertisements have been gray scale advertisements. In 1997, AP AdSEND has begun to distribute color advertisements as well as gray scale advertisements, using the ability of the PDF format to transmit color, including device independent color.

In addition, newspapers that receive PDF advertisements can use the "Export to EPS" feature of Acrobat Exchange to incorporate an EPS version of the PDF ad into a page layout. This capability eliminates the need for many of the manual production processes where errors are likely to occur. Combining PDF advertisements directly in a page layout can result in a substantial cost savings to newspapers, because doing so can eliminate errors and the need to offer rebates or make good on misrun advertisements.

Summary

CGATS.12 is really two standards efforts that are closely related. CGATS.12-1 is a standard for the exchange of complete PDF files. A complete file is one in which all the visual elements and all the resources needed to present those visual elements are included in the file. This means embedding all fonts used, the high resolution data for all replacement images and adjusting the color data for the intended printing conditions. CGATS.12-2 is a standard for "incomplete" exchange in which some specific elements and resources may be omitted from the exchange by prior agreement between the participants in the exchange. For example, the high resolution images used in image replacement may already be at the printer when a PDF-X file is sent to him. In this case, there is no need to embed the high resolution data in the file as would be required for a complete exchange.

A draft of CGATS.12 -1 is being prepared for ballot by the CGATS SC 6 subcommittee during the summer of 1996. This proposed standard specifies how to use PDF for the exchange of complete digital prepress data files among graphics arts establishments. This standard references the PDF 1.2 specification with a small number of extensions. These extensions include indications of whether the PDF file data has been trapped or not, what CMYK color space the data was prepared for and a mechanism that allows the high resolution replacement images to be transmitted with the PDF file. The standard specifies the use of PDF with only a few restrictions that are intended to insure successful exchange. These are primarily the use of CMYK color spaces and the limitation of compression to lossless compress techniques, such as Flate and RunLength.

Assuming normal progression of the standard and no new problems, CGATS.12-1 should be an approved ANSI standard around May, 1998. This document has also been submitted as a new work item to the ISO TC 130 Committee which is the international graphics arts standards organization. It seems likely that TC 130 will accept this work item and progress PDF-X in the international arena on a similar, but slightly delayed time scale.

Selected Bibliography

Adobe Systems Incorporated

1990 *PostScript Language Reference Manual, Second Edition,* Addison-Wesley, ISBN 0-201-10174-2, 1990.

Bienz, T., Cohn, R. and Meehan, J. R.

1997 *Portable Document Format Reference Manual, Version 1.2,* Adobe Systems Incorporated, San Jose, CA, (November 1996)

ISO

1997 ISO 12639 *Graphic technology - Prepress digital data exchange - Tag image file format for image technology (TIFF/IT)*, International Standards Organization, Geneva, 1997

McDowell, David

1997 "Status of TIFF/IT", *TAGA Proceedings*, Technical Association of the Graphic Arts, Rochester, NY.