PDF及其在电子出版领域的应用

杨道良　常　明　任晓霞

1　PDF概述

PDF(Portable Document format)是一种结构化的文档格式。它由美国著名排版与图像处理软件公司Adobe公司于1993年首次发布(1.0版),Adobe公司同年推出了相应的支持软件产品系列Adobe Acrobat 1.0;随后Adobe公司又对它进行修订和升级,于1994年发布了1.1版,并推出了支持软件产品系列Adobe Acrobat 2.0及2.1版。PDF最新版本1.2版于1996年11月27日发布,相应支持软件产品系列Adobe Acrobat 也升级到3.0版。1997年底国际标准化组织已经开始酝酿将PDF接纳为国际标准。

1.1　PDF与PS的比较

页面描述语言PS(Postscript)也是由Adobe公司拥有的一项事实上的印刷工业标准。它能描述精美的版面,在目前的印刷领域仍占据统治地位。PDF从PS发展而来,在对页面的描述方面它们有几乎相同能力和相似的描述方法。PDF采用与PS相同的着色模型(Imaging Mode)来表现文字和图形。与PS语言一样,PDF的页面描述指令是通过将选定的区域着色来绘制页面的。着色的区域可以是字母轮 廓、直线和曲线定义的区域以及位图;着色的颜色可以是任意的;页面上的任何图形都可以被裁剪成其他形状;页面开始时是全空的,各种指令将不同的图形绘制到页面上,新的图形是不透明的且可以覆盖旧的图形。

虽然如此,PDF与PS相比,还是有很大不同。主要表现在以下几方面:

PDF文件中可以包含交互对象如超链接、交互表单等。而PS没有。

PDF是一种文件结构,而PS是一种编程语言。因此PDF具有比PS更高的处理效率。

PDF的严格结构定义允许应用程序对其中的对象进行随机存取,而PS只能顺序存取。例如要访问一个PS文件中的第100页,必须先顺序解释其前99页后,才能找到第100页,而在PDF中对每一页的存取都是一样快的。

PDF中包含有字库的规格尺寸等字库描述信息,以便在字库不存在时进行字库仿真(而非简单的字库替代),保证文档显示的一致性。

1.2　PDF与html的比较

html是SGML(Standard Generalized Markup Language)的一个应用,是目前internet上主要的信息发布形式。它可以描述出web页面基本的样式,图文并茂,并有交互及超连接功能,配合Java或script能有一些处理能力,还可以通过cgi与服务器交互。PDF同html一样也具有表单交互和超级链接功能,适合于网上发布信息。但与html不同的是PDF还具有描述精美版面的能力。PDF实现了纸张印刷和电子出版的统一。排版后的内容保存成PDF文件,则能在交付印刷的同时,进行网络发行(需增加适当的交互内容),而不必象目前的一些作法一样,需要两组人员,一组为纸张印刷进行排版生成PS,另一组为电子出版创作html文件,造成资源和人力浪费,生产效率低下。

html除了没有版面描述能力外,还经常出现信息的不一致性(如不同平台,不同浏览器,不同尺寸的浏览器窗口看到的同一web页面呈现出不同样子)。而在PDF中已经很好地解决了这个问题。

1.3　PDF的特点

PDF的特点归纳如下:

可传递性。PDF文件支持7位Ascii码和二进制两种编码方式,可以正确地在各种网络环境下传输。

平台无关性。PDF文件具有软、硬件平台独立性。用户在不同的环境下(如不同语言的操作系统、不同的硬件平台)看到的PDF文件的版式和内容都与作者创作完成时的情况完全一致。这个特点非常适合于信息交换,免除乱码的苦恼。

字体无关性。PDF文件中可以自带字体或字体描述信息,在用户的系统中缺乏所需字体的情况下,仍然能正确显示。

支持多种压缩、编码方式,文件更紧凑。压缩、编码方式有:Asciihex、 scii85、lzw、runLength、ccitt group3、ccitt group 4、jpeg、flate。

支持交互操作。可包含交互表单和超链接。支持声音、动画。

支持对页面的随机存取。

支持不断追加的修改方式,便于少量修改、提高效率。

安全性控制。支持各种不同级别的安全性,如只能阅读,不能打印和选择文字;可阅读、可打印,但不能修改;可阅读、可打印、可修改等。这种安全性控制对保护电子出版物的版权非常重要。

2　PDF的结构

2.1　PDF文件结构

PDF的文件结构(即物理结构)包括四个部分:文件头、文件体、交叉引用表和文件尾,参见图1。

文件头指明了该文件所遵从PDF规范的版本号,它出现在PDF文件的第一行。如%PDF-1.2 表示该文件格式符合PDF1.2规范。

文件体由一系列的PDF间接对象(inDirectob Ject)组成。这些间接对象构成了PDF文件的具体内容如字体、页面、图像等等。

交叉引用表则是为了能对间接对象进行随机存取而设立的一个间接对象地址索引表。

文件尾声明了交叉引用表的地址,指明文件体的根对象(catalong),还保存了加密等安全信息。

根据文件尾提供的信息,PDF的应用程序可以找到交叉引用表和整个PDF文件的根对象,从而控制整个PDF文件。

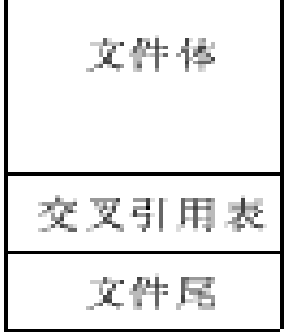


图1

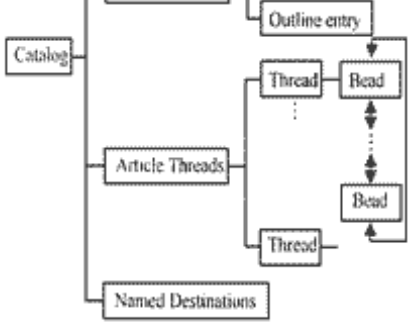


图2 PDF 文档结构图

2.2 PDF文档结构

PDF的文档结构是PDF文件内容的逻辑组织结构。它反映了文件体中间接对象间的等级层次关系。PDF的文档结构是一种树型结构,如图2。树的根节点就是PDF文件的根对象。根节点下有四个子树:页面树(Pages tree)、书签树(out line tree)、线索树(Articleth reads)、名字树(named Destination)。其中在页面树中,所有页面对象都在树的叶子节点,树中的子节点将继承父节点的各属性值作为相应属性的缺省值。书签树中则按树型层次等级关系将书签(booKmarK)组织起来。书签建立了书签名与一个具体页面上的位置的关联,它使得用户可以按书签名字来访问文档的内容。由于书签可以有层次,能用来组织文档的目录,所以有时又将书签树称作目录树。线索树则将文章线索及线索下的文章块(Article bead)按树型结构组织起来进行管理。文章块是预定义好的一个页面上的区域,它一般是读者感兴趣的一段文字或图像,它的目的是让整个可视区只显示这个特定区域而避免页面其他部分的干扰。文章线索将预定义好的文章块串接起来,如果读者按文章线索进行阅读,则浏览器只按顺序显示该线索中的各文章块,从而使读者只读自己感兴趣的内容,而不必按顺序阅读。至于名字树则是建立了一种字符串(名字)和页面区域的对应关系,树中的叶子节点保存字符串及对应的页面区域,而非叶子节点只是一种索引,以便让应用程序能快速存取到叶子节点。名字树的作用就是让PDF文件中的其他对象能够用字符串名字来代表一个页面区域。

2.3 PDF中的资源

PDF中的页面内容(如文字、图形、图像)都保存在页面对象的contents关键字对应的流对象(下面简称内容流)中。内容流中用到了很多基本对象如数字、字符串,这些都是用直接对象表示的。但还有其他一些对象如字体,本身就是用字典对象(Dictionary)或流对象(stream)来表示的,无法用直接对象表示,而内容流中又不能出现任何间接对象(否则无法与内容本身的数据区分),于是就将这些对象命名,并在内容流中用相应的名字来表示它们。这些用名字来表示的对象就称作命名资源(named resources)。

在页面对象中有一个资源项(resource skey),该项列出了内容流中用到的所有资源,并建立了一个资源名字与资源对象本身的映射表。

PDF中的命名资源有:指令集(Procset)、字体(font)、色彩空间(color space)、外部对象(xobJect<包括image、form和Pssegment>)、扩展的图形状态(extended graphics state)、底纹(Pattern)、用户扩展标记列表(Property list)。

非命名资源有:encoding、font Descriptor、halftone、fuction、 CMAP。由于非命名资源都是被隐含引用的,因此没有命名的需要。

2.4 PDF页面描述指令

PDF一共有60个页面描述指令。这60个页面描述指令描述了页面上的一系列图形对象。这些图形对象可分为四类:路径对象(Path object),文本对象(text object),图像对象(image object),外部对象,参见图3。它是构成所有页面的基本元素。

3 PDF文件的生成

目前PDF的生成有两种途径:

· 通过打印的的方式生成PDF,就是通过一个虚拟的PDF打印机将应用程序的文字和图形指令(如windows下的gdi指令或Mac下的QuicK-Draw指令)转换为PDF指令并保存在PDF文件中,参见图4。在安装了Adobe Acrobat PDF writer之后,从理论上说所有的具有打印功能的应用程序都能将待打 印的内容打印到PDF文件中。但目前生成中文PDF文件尚有很多问题。

图3　页面上的图形对象

图4　通过打印方式生成PDF

由PS转换到PDF是另一种生成PDF的方法,它是由应用程序先将待打印的内容发排到PS文件,再由Adobe Acrobat Distiller将PS文件转换成PDF文件,参见图5。

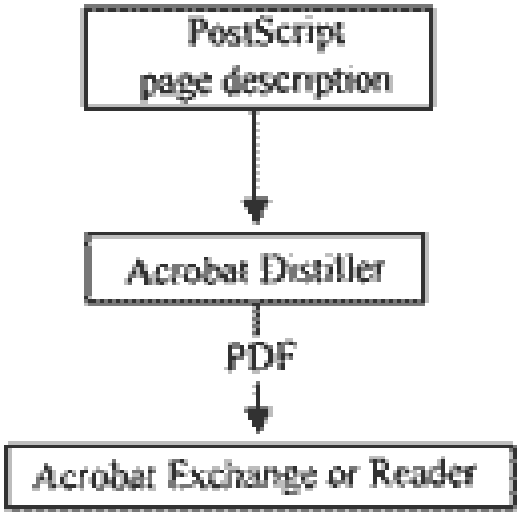
两种生成PDF的方法各有利弊。通过打印方式生成PDF的优点是和应用程序能够紧密结合,在用户看来是从应用程序直接生成PDF,但缺点是由于gdi指令集和QuicK-Draw指令集本身的局限,难以生成高精度的PDF。而从PS转换到PDF虽然多了一道工序,但由于PS本身具有高精度的描 述能力,因此生成的PDF可以达到印刷级的质量和精度。

生成PDF文件之后,用户就可以用Acrobat exchange或reader来阅读和打印。还可以使用Acrobat exchange给PDF文件增加如页面缩像、超链接、书签(或目录)、注释等交互属性。

采用Adobe提供的工具生成PDF目前都存在中文支持方面的问题。如不支持中文字库的下载,中文显示依赖操作系统等等。著名的中文激光照排系统开发商北大方正集团近期将推出全面支持中文的PDF生成工具。

4　PDF在电子出版领域的应用

由于PDF具有诸多适合电子出版的特性,目前在电子出版领域的应用日益增多。具体应用可分为三种情况:制作CD-ROM电子出版物,与html混合使用发布信息,独立采用PDF制作主页及发布信息。



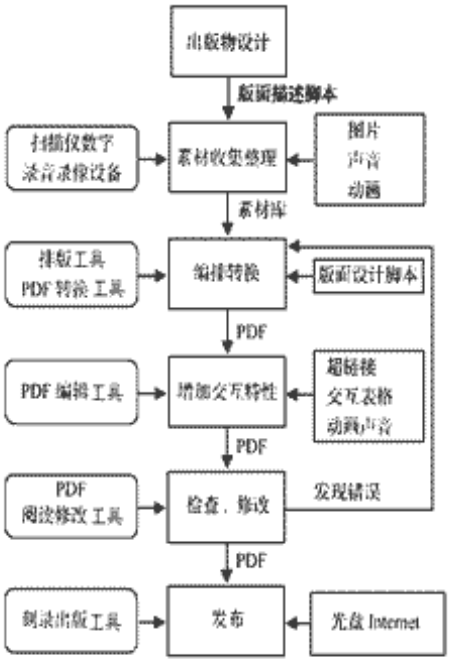


图5　从PS转换生成PDF

图6　用PDF进行电子出版的流程

不管是哪种情况,采用PDF进行电子出版都可能经历下面几个步骤(参见图6):

1)出版物的设计。根据未来读者的机器类型配置、以及采用的信息发布介质来设计出版物。

2)素材收集。根据设计收集整理所需的各种素材。

3)编排与转换。采用自己最熟悉和喜欢的工具完成如绘图、扫描、排版等工作,并按第三部分所述的方法将结果转换成PDF。

4)给PDF增加交互特性。即用Acrobat exchange或类似的工具给PDF文件增加如超链接、按钮、交互表格、动画(Movie)、声音等特性。

5)检查修改。对初步成型的出版物进行预览和检测,如发现少量错误(如错别字等),则可直接在Acrobat exchange或类似的工具中进行修改,如有较大错误,则需回到第3)步,进行重新编排。

6)发布。将成品PDF文件刻录到CD-ROM或放到www站点进行发布。用PDF制作CD-ROM出版物是目前应用最多的情况,在国内也有大量实例。如广为流传的《黄金书屋》光盘及中国大百科全书出版社出版的《中国大百科全书·光盘版》都是采用PDF进行光盘出版物制作,是较为成功的例子。

由于现在只有少量的www服务器支持PDF,因此采用PDF独立进行主页设计和信息发布在未来的一段时间内还不太现实。但已有大量www站点采用html和PDF混合的方式进行信息发布,如在html框架中嵌入PDF,两者可以无缝结合。对于支持PDF的www站点(如Adobe的www站点),用户从上面阅读PDF和html是等效的,可以边传边读(用户浏览器中要事先安装阅读PDF的插件或控件)。而从不支持PDF的www站点上阅读PDF,则只能等PDF文件完全下载到本地之后,用户才能阅读。目前已有大量的电子杂志(或杂志的电子版)采用PDF在internet上发行(如美国《科学》杂志的电子版),internet上的PDF资源也越来越多,一些著名internet搜索引擎也逐步开始支持PDF的搜索。可以预见,随着更多的www站点对PDF的支持,PDF将在未来的internet上逐步占据主导地位。

PDF的出现不仅对电子出版带来巨大影响,也对传统的印刷流程产生了冲击。传统的以PS为中心的印刷将面临PDF的挑战,未来的PDF RIP(raster image Processor)将会逐步取代PS RIP,从而真正实现一次制作、多次使用(印刷、网络发行)的思想。

参考文献

[1]　Adobe systems IncorPorated. Portable Document format Reference mannul Version1. 2, 1996,

11
［2］ Adobe systems IncorPorated. Postscript Language Reference mannual second edition[3] utK. wolf. PDF：Can it Move from wish List to Prodution workflow?. seybol Dreporton Publishing system. Volume27，nuber15