一种在电子出版中融合固定版面与流式信息的方法

仇睿恒[1,2],汤 帜[3]

(1.中关村科技园区海淀园博士后工作站北大方正集团有限公司分站,北京 100871;
2.数字出版技术国家重点实验室(筹),北京 100081;3.北京大学计算机科学技术研究所,北京 100871)

摘 要: 随着硬件条件的提高和网络技术的发展,特别移动终端的快速发展,电子文档的使用环境日趋多样化,但相关技术却面临着更大的挑战.这是因为固定版面与流式信息之间存在本质的矛盾,难以进行融合、协同工作.虽然人们尝试了一些方法来解决这个问题,但是效果都不甚理想.我们在研究现有技术的基础上,提出了一种新的基于版面块的文档模型,并赋予其固定版面的特性与必要的流式信息,以适应多样化的终端环境,能够解决电子文档出版中的固定版面与流式信息融合的问题.实现效果说明,本文提出的文档模型在实际使用中具有很大的潜力.

关键词: 文档处理; 固定版式; 流式文档; 电子出版

中图分类号: TP317.1 **文献标识码:** A **文章编号:** 0372-2112 (2012) 11-2276-06

电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.11.021

A Novel Method of Merging Fixed Layout and Fluid Information for Electronic Publishing

QIU Rui-heng[1,2], TANG Zhi[3]

(1. *Postdoctoral Substation of Peking University Founder Group, Postdoctoral Workstation of the Zhongguancun Haidian Science Park, Beijing* 100871;
2. *State Key Laboratory of Digital Publishing Technology, Beijing* 100081;
3. *Institute of computer science and technology, Peking University, Beijing* 100871)

Abstract: With the development of hardware and network technology, electronic publishing is widely used nowadays on Internet. However there are many difficulties to render electronic documents properly for various reading devices, because it is hard to merge fixed layout and fluid information together due to the significant difference between them. Some methods have been proposed to solve this problem, but the final results are not ideal. In this paper we proposed a novel document model based on page block, and fixed layout and fluid information are both constructed on it. This method can merge fixed layout and fluid information together for electronic publishing and adapt to various reading devices. The experimental results show that the document model proposed in this paper performances well in practice.

Key words: document processing; fixed layout; fluid information; electronic publishing

1 引言

文档作为信息的载体,在人类历史和社会进步中发挥着重要作用.随着电子技术的发展,电子文档日益普及,而其用途也日益多样化.例如:“标文通”[1]、OOXML[2]等被广泛适用于办公领域;PDF[3]、XPS[4]则大量使用于档案、电子公文和数字出版等领域;而 ePub[5]则适用于手机移动领域.近年来,电子文档格式作为内容的载体和阅读技术的容器越来越受到各方的重视,各机构、厂商都在积极推动其文档格式标准工作.随着相关工作的进行,人们发现如何在固定版式文档与流式文档间共享、交换信息乃至相互融合是一个亟待解决的关键问题.因此,本文将在电子文档出版的背景之下对这一问题进行分析与探讨.

2 固定版式文档与流式文档

根据版面计算方式,电子文档可以分为固定版式文档与流式文档.固定版式文档是一种以保存文档显示或打印结果为目的的文档,采用图元为基本单位,对文字、图形、图像等信息进行精确定位描述,从而在不同的输出环境下都能获得一致的绘制结果.固定版式文档一般由一组页集合构成.每个页包含了一组基于精确空间位置的图元定义、组合和引用,从而构成版面的描述.而流式文档则侧重于记录作者在编辑文档时的意图,基于文档内容流、逻辑结构和样式,并利用各种排版算法计算出最终的显示结果.流式文档一般由章节、段落、节、

收稿日期:2010-10-18;修回日期:2012-07-25
基金项目:国家 973 重点基础研究发展计划(No.2012CB724108)

样式等要素构成,按照作者的编辑意图组织成树状结构.二者的根本区别在于版面绘制结果的确定时机:固定版式文档在生成时就确定了显示结果,而流式文档则在打开时动态的计算绘制结果.

在实际使用中,文档打开的环境千差万别,同时文档软件的排版算法也在不断修改.这就造成流式文档在不同的平台上的绘制结果很难保持一致,甚至会有很大的不同,从而产生的了"跑版"的现象.这对于很多注重绘制结果的应用是致命的缺陷.而固定版式文档则着重于精确定位描述,对文档的章节、段落、表格等等逻辑信息并不关心,这使得再次编辑这类文档时会面临较大难度.

3　电子文档出版中的文档格式

电子文档出版是一个复杂的过程,往往会经过写作、排版设计、校对、发布等阶段.人们在不同的阶段会关注文档中不同的信息.例如,写作时作者会专注于创作文档的内容,并添加适当的说明章节等逻辑结构信息;而在排版设计时,编辑则着重为文档的各部分内容赋予排版属性以将其安排到合适的版面中,同时调整文档的部分逻辑结构,修正词句等.

根据出版过程,我们可以将一份文档的信息分为四个部分:内容、逻辑结构、排版属性、绘制.随着文档出版过程的推进,人们会关注于其中不同的部分,版面越来越固定,从而获得最终的出版书籍.由于流式文档与固定版式文档关注于文档中不同部分的信息,因此在使用电子技术制作、出版、阅读书籍时,经常会需要固定文档格式与流式文档格式协同工作.多种文档格式的协同工作意味着在不同文档格式间能够互相进行信息同步和数据转换,例如:当发现小样中的内容错误时,需要迅速定位到原始内容中的位置;完成排版后,需要生成相应的固定版式文档以供印刷使用;当最终出版的文档到达最终读者时,读者也希望能够从其中提取出诸如实验数据表格这样的逻辑结构信息.这些困难都对固定版式文档与流式文档的信息交互、转换与融合提出了要求.

随着移动阅读的普及,丰富的表现形式、多样化的交互手段以及特征各异的移动终端,使得人们需要付出更多的精力才能制作出一本高质量的电子文档.特别是,当一份文档需要在具有不同显示特征的终端上展示时,制作者需要付出大量的额外工作以制作多份不同的电子文档版本来应对不同的移动终端.这大大提高了电子文档制作的成本,增加了发布服务的复杂性和服务器的资源开销.因此电子文档的一次制作、多平台多次利用也成为了电子文档出版对于固定版式与流式信息融合的迫切需求.

4　融合固定版式文档与流式文档的尝试

如前文所述,当前的电子文档出版对固定版式文档与流式文档的融合有着迫切的需求.但二者的融合却一直没有实现,因此相关领域的研究人员一直在对其进行研究,希望能够改变这一现状.

从一般的观念来看,版式意味着文档内容与显示的固定性,而流式则象征着其可变性.但是,Levy[6]在1994年就对文档中的固定性与可变性进行研究和总结.他认为,虽然传统文档的关键特征是用固定的符号在稳定的载体上以固定的表现形式记录固定内容,而随着信息电子化,交互、可修改、可重排等可变性特点则成为了电子文档的重要特点,但是这些并不意味着固定性和可变性是对立的.无论是传统文档还是电子文档,固定性和可变性总是同时存在的,只是在表现形式与程度上有所区别.这就说明了固定版式文档与流式文档的融合存在可能性.

而在 Dori[7]在其书中提出了文档的逻辑结构与物理结构的概念,并建立相应的文档模型.在这个模型中,物理结构指的是文档的页、块等与版面显示、布局相关的信息,而逻辑结构则关注于文档的章节、标题、段落等信息,二者共用文档中的数据,如文字、图像等.

基于类似的文档模型,Adobe 在 1999 年推出的 PDF 1.3 规范中引入了 logical structure,尝试在固定版式文档中加入流式信息,并且在 2001 推出的 PDF 1.4 中引入了 tagged PDF 来完善流式信息的表达.之后 adobe 又在其 MARS[8]文档格式中使用 XML 对这部分信息进行结构化的描述.

基于 tagged PDF 技术,Hardy[9,10]等人提出一种利用 XML 模板生成 tagged PDF 文档的方法.但该方法对对 PDF 文档的内容有较高的要求,实用中存在一定的困难.而李宁[11]则针对"标文通"与 Tagged PDF 的信息交换进行了实验,为减少办公文档的跑版问题提供了积极的借鉴意义.

此外,Bloechle 等人也基于 Dori 模型开展了一系列的研究工作.Bloechle 在 2006 年提出了 XCDF[12]格式,通过引用版面中的文字、图像、图形等内容来构造一个包含完整版面信息与流式信息的文档.与 Tagged PDF 相比,XCDF 没有类似 PDF 的历史包袱,其中版面信息与流式信息的结合更为紧密合理,并且采用了 XML 来描述相关信息,使得其构造、使用更为方便.2008 年,Bloechle 在 XCDF 的基础上提出了一种从已有固定版式文档中重新构造文档逻辑结构的方法——Dolores[13].之后,Bloechle 又对 XCDF 格式进行了优化[1],大大缩小了所生成的文档体积,使其更利于使用.

在尝试采用 tagged PDF 等技术来为固定页面添加

流式信息的同时,人们也尝试了其他一些方法以达到避免跑版、方便数据交互的目的.这些方法包括为流式文档设定统一版面算法、增加多种参考标记,来辅助流式文档固定版面等.例如 ePub 采用 XML 与 CSS 的技术解决方案,通过数据与表现分离、丰富的样式描述等方法希望在不同环境下都能获得相似的显示效果.而"标文通"则通过分离内容和样式来解决不同应用间的数据交换问题[14].但这些技术都没有从根本上解决固定版式文档与流式文档的融合问题.

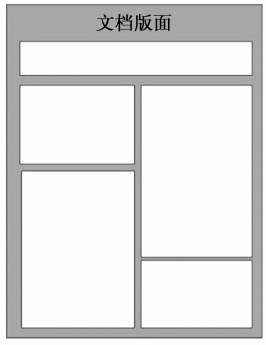
5 一种融合固定版面与流式信息的方法

尽管人们在融合固定版式文档与流式文档上已经进行了很多工作,但是目前仍然没有一种文档格式可以在电子文档出版领域取得很好的效果.例如:ePub 在避免跑版、版面表现力等方面仍有欠缺,所以 IDPF 正在对这些缺陷进行研究,希望在 ePub 的下个版本中引入新的机制解决这些问题[1];而 Tagged PDF 在实际使用中往往会遇到两个困难:

(1)剪裁的使用.在固定版式文档中,剪裁被看作为一种限定输出区域的操作.多个剪裁操作可以相会作用,并影响后续的绘制内容.但是其在流式信息中很难被表述、使用.

(2)复杂效果的描述.固定版式文档中存在大量拼接的图元、相互叠加的图层.这些数据并不能直接用于构造文档的逻辑结构,而需要根据空间关系进行拼接、叠加等操作后才能确定其所要展示的内容.

经过比较 Dori 的文档模型与电子出版时文档编排版面的过程,我们发现这两个困难是由固定版面与流式信息的不同出发点引起的.Dori 的文档模型假设版式信息与流式信息的地位是对等的.但在显示制作、编辑文档的过程时,由于版式信息与流式信息之间存在的同步问题,只能基于固定版面方式或流式方式中的一种.因此,在电子文档中保存的数据总是会倾向于固定版式信息或者流式数据中的一种.例如 tagged PDF 中的流式信息就是以固定版式数据为基础的.因此,尽管 tagged PDF 保存了文档的逻辑结构,但是其逻辑结构基础的并非 Dori 模型中内容数据,而是固定版式数据内容流,所以在 tagged PDF 中固定版式信息与流式信息并没有达到真正的融合,进而造成了其在实际使用存在一些不足,例如不同内容流之间的整合、富样式内容的重排等.



5.1 基础文档模型

针对现有方案中的不同,我们完善了以往工作[1],提出了一种新的针对电子文档出版的文档模型(下文简称为 CEBX),融合了必要的固定版式数据与流式信息,以达到一次制作、多平台多次利用的目的.与 tagged PDF 不同,我们以版面块为基础构造 CEBX 文档模型,将版面块作为固定版式信息与流式信息融合的基本单位,从固定版式数据出发构造相应的流式信息,并根据需要赋予其必要的交互特性,以达到电子文档出版的需要.

如上文所述,文档中的信息总是会倾向于版式与流式信息中的一方.对于电子文档出版来说,一个文档会面临多样化的展示环境,因此流式信息与版式信息同样重要.但是考虑到文档出版对版面的不变性与表现力有很高的要求,而流式文档由于其自身特点限制,难以杜绝跑版现象的发生.因此与 tagged PDF 一致,CEBX 利用固定版式数据作为基础来构造文档的流式信息,以从根本上解决跑版问题.

电子出版中文档在出版发布后并不需要进行再次的修改与编辑,因此我们的工作主要关注于文档的排版制作、发布后的多平台展示以及流式信息提取,即在终端解决版式信息与流式信息的融合问题.而在办公等领域中用户在生成文档后,往往还需要对其进行多次编辑,即在制作端对固定版面和流式信息进行兼容.这是电子出版流程与办公文档编辑流程的本质区别,这也促使我们从文档的版面制作出发来寻找解决问题的方法.

图 1 展示了 CEBX 文档模型.其核心意图是从固定版式数据出发,将一个版面根据制作者的意图分为多个独立的版面块,各版面块由固定版式中的图元构成并具有一定的独立含义.将这些版面块按照空间关系进行组合便可得到所需要的固定版式页面;而如果将他们按照一定逻辑结构进行组织则可以得到文档的流式信息.与现有方法相比,CEBX 文档模型认为在文档编辑、排版的过程中最基本的单元并不是数据,而是每页中具有一定基本语义的版面块.这个模型结构是由三个因素决定的:

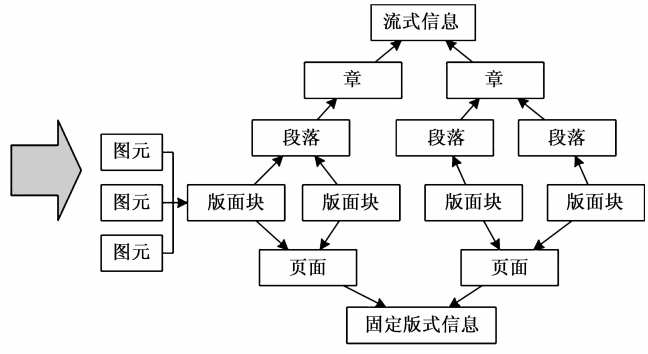


图1 基础文档模型

(1)人们在阅读某一个文档页面时,会在视觉上将这个页面分成多个区域,然后从中选择合适的区域按照合适的顺序进行阅读.

(2)在制作一个文档版面时,人们会将版面划分为多个块或者根据相应的页面划分规则自动得到多个块,例如页眉、页脚、标题、分栏等,然后将数据按照一定的排版方式填充在这些块上.

(3)版式信息与流式信息的相似点.通过观察页面中局部区域的版式数据,我们发现其描述与相应的流式数据有很高的相似性,包括文字编码、字体描述、颜色空间、图片、变换缩放等等,甚至二者的数据顺序也有高度的一致性.因此,基于版面块级别来构造文档模型可以最大程度的共享基础数据的定义、减小描述数据量、减少版式信息与流式信息间的同步代价.

基于这三点因素,我们在 CEBX 文档模型中将页面中具有一定语义的版面块作为基本单元,并在利用这些版面块构建整个文档.事实上,很多文档智能识别系统也是根据人类在阅读文档的视觉行为,先将待识别的页面进行分割,再分块进行识别,最后重新组织得到最终的识别结果.这也是采用版面块作为文档模型基础单元合理性的一个佐证.

5.2　CEBX 文档结构

在确定了基础文档模型后,之后的工作便是在其基础上构造一个具有完整版式信息和流式信息的文档.其中,版式信息由文档中所包含的页面以及页面间的关系(页树)、构成页面的版面块(Page block)组成;而流式信息则包括了标题、段落、节、句等信息.

我们将版面块进一步拆解和细化,得到一系列基本图元,如文字、图像、线条、渐变等等,并将这些图元组织在一系列资源文件中.固定页面则被划分为多个版面块,每个版面块由一系列图元构成.这样就完成了固定页面的拆分和版面块的声明,而流式信息则会构造在这些版面块之上.

除了以上所提及的信息外,链接、声音、视频、显示变换、数据收集提交(电子表单)等简单的交互要素则附着在各图元上,在版式信息与流式信息中共同使用,而区域动画效果(例如页面局部放大等)等则会在相应页面区域中进行描述.

图 2 是一个 CEBX 文档的结构示例.其中资源 PageRes_1 定义了一个文字图元“Peking University”,并指定了其所使用的字体、大小、相对位置等样式信息,然后在固定页面 Page 的版面块 Block100 中进行展示,而在 structure root 中则通过对该版面块的引用完成了流式信息的构建.

5.3　剪裁

剪裁对于融合版式数据与流式信息是一个巨大的障碍,所以在 tagged PDF 的流式信息中剪裁操作被直接忽略了.而在 CEBX 中则采用了另一种处理剪裁的方

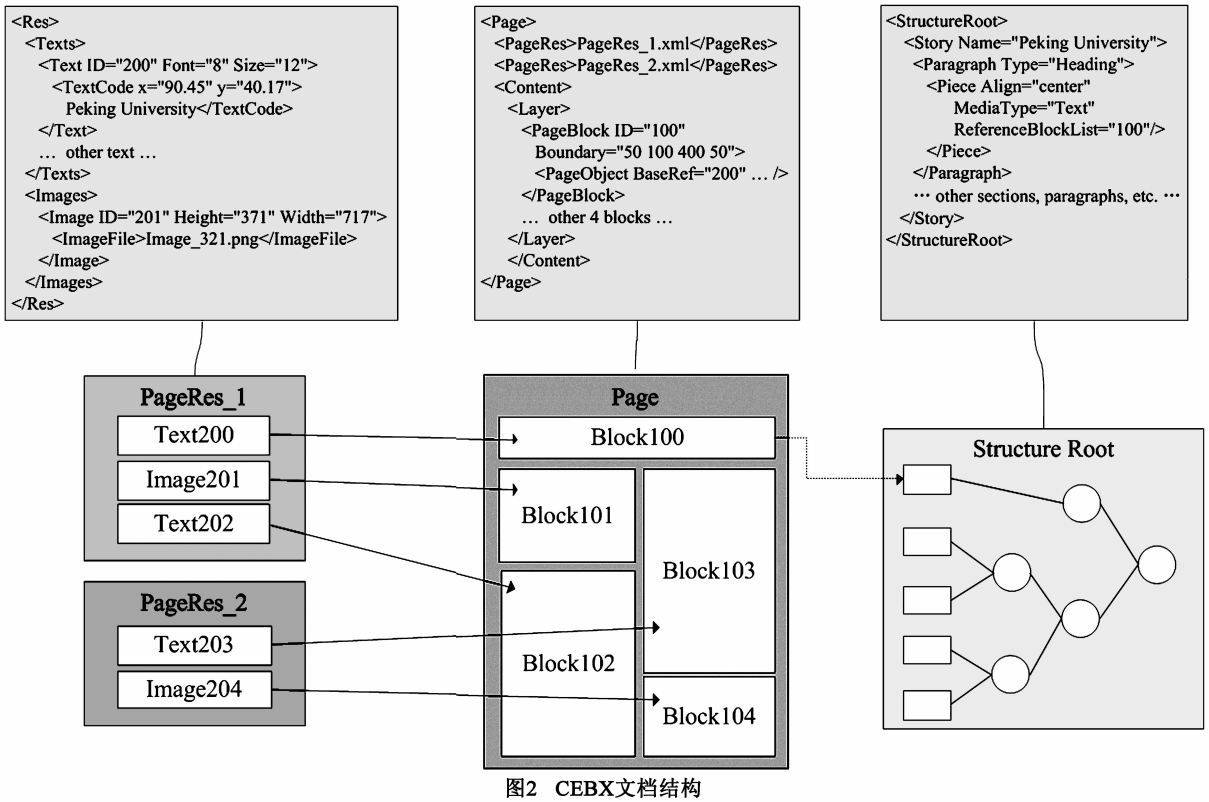


图2　CEBX文档结构

法.当版式文档中出现一个剪裁操作时,制作者的意图往往是将被剪裁影响的图元在视觉上进行叠加、组合、修剪以达到一个特殊的效果.因此,我们在 CEBX 的文档模型中将剪裁的概念进行了改进,认为剪裁是图元的一个属性,用于设定该图元的有效绘制区域.剪裁属性只会影响该图元的绘制效果,而不会影响后续出现的图元的绘制.这就使得每个图元在单独绘制时不需要依赖其前后图元,有了完备的自描述性,可以直接在流式状态下使用.而当多个图元相互叠接通过剪裁达到某种特殊效果时,我们则会将这些受到同一剪裁区影响的图元组合成复合图元(一组基础图元的容器),并赋予其剪裁属性,以达到所需的效果.这种方式一方面切合文档模型中版面块的概念,被剪裁的图元形成了一个具有独特含义块,可以直接在流式应用中使用而不破坏显示效果;另一方面这种将剪裁作为图元属性的方式也经常使用在流式文档中(例如"标文通"中图像就具有类似的属性定义),从而达到了图元级别上的信息共享.

5.4 实现效果

为了检验所设计的文档模型在实际使用中的效果,我们在对 CEBX 的制作、版式阅读以及流式重排的流程进行了实现.图 3 展示一组利用 CEBX 文档模型融合固定版式与流式信息的实例,并将效果其与 tagged PDF 进行对比.图中所选取的实例是来自于朱玉峻[18]等人发表的论文,其中包含了一些由多个图形叠加形成的公式.图 3 的上方是该文档原始版面,左侧是利用 PDF 技术得到的版面重排结果,右侧则是利用 CEBX 文档模型所得到的重排效果.

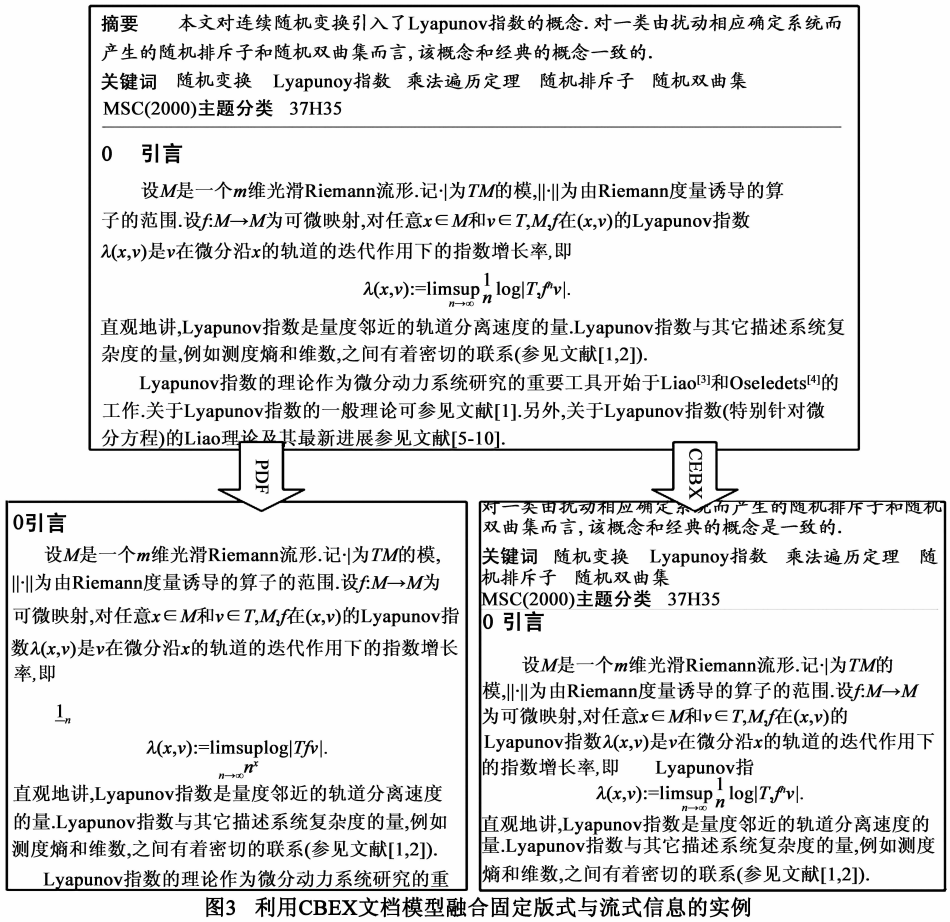


图3 利用CBEX文档模型融合固定版式与流式信息的实例

利用 PDF 技术得到的版面重排结果中主要存在几个方面的问题:

(1)局部的顺序错误.这是由于 tagged PDF 基于内容来构造逻辑信息,对于分属于不同流之间的内容难以整合到一个统一的顺序中.

(2)公式错误.这是由于其在设计时没有考虑到局部版面的相对固定的特性.这对于公式、复杂图形等图元的重排是非常重要的.从示例可以看到,虽然在 CEBX 文档模型中并未定义公式的描述,但通过对版面块的引用也达到了很好的重排效果.

(3)样式展示.CEBX 在版面块上进行样式的统一描述,而不是通过版面和流式信息分别进行描述,因此在版面重排时,CEBX 能够达到一致的样式效果.

从这组实例的展示和分析结果可以看到,CEBX 文

档模型与现有技术相比在实际使用中具有很大潜力,能够同时获得很好的固定排版和流式重排效果.

6 结论

综上所述,为了解决电子文档出版中同一文档在不同显示环境下的阅读问题以及对版式文档进行流式信息提取的问题,需要一种融合固定版式与流式信息的解决方案.相关的研究人员一直就这个问题进行探索与尝试.我们对他人的解决方案进行了研究,并结合了文档排版的特征,以固定版式数据为基础,利用版面块来构造文档的固定版面和流式信息,并赋予其必要的交互要素,以达到一次出版、多平台多次利用的目的.

由于我们的工作是从终端出发解决版式文档与流式文档的融合问题,并未涉及文档发布后的后续编辑与修改,因此仍然需要对本文所提出的方法在办公文档领域中的应用进行论证.此外,本文方法中所给出的流式信息是一个较为简单的结构,在以后工作中需要进一步借鉴"标文通"、OOXML 等工作小组的研究成果对其进行丰富、细化.

参考文献

[1] GB/T 20916-2007.中文办公软件文档格式规范[S].

[2] ISO/IEC 29500:2008, Information technology—Office Open XML file formats[S].

[3] ISO 32000-1:2008, Document management-Portable document format—Part 1:PDF 1.7[S].

[4] Microsoft Corporation. XPS Specification and Reference Guide [EB/OL]. http://www. microsoft. com/whdc/xps/default. mspx.

[5] Renear A and Salo D. Electronic Books & the Open Ebook Publication Structure[M]. The Columbia Guide to Digital Publishing, 2003, 455 – 520.

[6] Levy D M. Fixed or fluid?: document stability and new media [A]. Proc. the 1994 ACM European Conference on Hypermedia Technology[C]. New York: ACM Press, 1994. 24 – 31.

[7] Dori D. The Representation of Document Structure: a Generic Object-Process Analysis[A]. Handbook on Optical Character Recognition and Document Image Analysis[C]. Singapore: World Scientific, 1995. 421 – 456.

[8] Hardy M R B. The Mars project: PDF in XML[A]. Proc. the 2007 ACM Symposium on Document Engineering[C]. New York: ACM Press, 2007. 161 – 170.

[9] Hardy M R B, Brailsford D F. Mapping and displaying structural transformations between XML and PD [A]. Proc. the 2002 ACM Symposium on Document Engineering[C]. New York: ACM Press, 2002. 95 – 102.

[10] Hardy M R B, Brailsford D F, Thomas PL. Creating structured PDF files using XML templates[A]. Proc. the 2004 ACM Symposium on Document Engineering[C]. New York: ACM Press, 2004. 99 – 108.

[11] 李宁,田英爱,侯霞,梁琦.办公文档与固定版式文档格式关系探讨[J].电子学报,2008,36(B12):128 – 132.
Li Ning, Tian Ying-ai, Hou Xia, Liang Qi. A Discussion on relationship between revisable and non-revisable document formats [J]. Acta Electronica Sinica, 2008, 36(B12): 128 – 132. (in Chinese)

[12] Bloechle J-L, Rigamonti M, Hadjar K, et al. Xcdf: A canonical and structured document format[A]. Proc the 7th International Workshop on Document Analysis Systems[C]. New York: Springer, 2006. 141 – 152.

[13] Bloechle J-L, Pugin C, Ingold R. Dolores: An Interactive and Class-Free Approach for Document Logical Restructuring [A]. Proc. the 8th International Workshop on Document Analysis Systems[C]. Los Alamitos: IEEE Computer Society, 2008. 644 – 652.

[14] Bloechle J-L, Lalanne D, Ingold R. OCD: An Optimized and Canonical Document Format[A]. Proc. the 10th International Conference on Document Analysis and Recognition[C]. Los Alamitos: IEEE Computer Society, 2009. 236 – 240.

[15] 李宁,牟永敏,董慧,方春燕.文档格式中"内容"与"表现"的分离与融合[J].电子学报,2007,35(2):375 – 378.
LI Ning, MU Yong-min, et al. Separation and combination of dontent and appearance in document format[J]. Acta Electronica Sinica, 2007, 35(2): 375 – 378. (in Chinese)

[16] IDPF. EPUB 2.1 Working Group Charter-DRAFT 0.8[EB/OL]. http://www.idpf.org/idpf_ groups/IDPF-EPUB-WG-Charter-4-6-2010.html, 2010-06-04.

[17] Ruiheng Qiu, Zhi Tang, Liangcai Gao, Yinyan Yu. A novel XML-based document format with printing quality for web publishing[A]. Proc SPIE Conference on Imaging and Printing in a Web 2.0 World, Vol. 7540, 75400J [C]. Berlingham: SPIE, 2010. 1 – 10.

[18] 朱玉峻,张金莲.连续随机变换的 Lyapunov 指数[J].中国科学(A 辑:数学),2009,39(5):555 – 566.

作者简介

仇睿恒　男,1982 年生于江苏泰州.北京大学计算机科学技术研究所博士.研究方向为文档处理.
E-mail: qiuruiheng@gmail.com

汤　帜　男,1965 年生于浙江台州.北京大学计算机科学技术研究所研究员,博士生导师.研究方向为文档处理、数字版权保护技术.