数码印刷中 XML 和 PDF 流程整合探讨

作者：马桃林、刘庆华

【内容提要】对于出版商和广告商来说，以前传统印刷和数码印刷使用的印刷媒介服务是相对独立的。而今，所有的印刷流程，尤其是数码印刷流程控制都与网络产生了密切的关系。印刷厂需要在网络上发布广告，提供服务信息，对目标市场进行宣传……

对于出版商和广告商来说，以前传统印刷和数码印刷使用的印刷媒介服务是相对独立的。而今，所有的印刷流程，尤其是数码印刷流程控制都与网络产生了密切的关系。印刷厂需要在网络上发布广告，提供服务信息，对目标市场进行宣传。数码印刷可以实现在客户家里、客户网络打印机上完成、远程印刷等业务。

用于跨媒体出版的文件无须再做进一步处理，即可被重新利用或输出到各种媒体中（如传统的印刷品、HTML 界面、光盘数据库、电子书等）。XML 和 PDF 都可以作为跨媒体出版的文件格式。PDF 侧重于描述文档的打印或用于印刷输出的格式；XML 侧重于描述信息的内容本身，最主要的应用是作为数据交换的中介。

PDF 和 XML

PDF 文件为可携式文件格式，现在已经成为跨平台的通用格式，也是网上电子杂志等最热门的传递方式。PDF 文件可通过 PageMaker、InDesign、Word、Fits 等排版软件创建，通过 Adobe Acrobat 等软件阅读和修改。它也可畅通无阻地在任意平台上显示与阅读。PDF 处理的对象可以是巨幅画片、复杂图纸、文字材料等，应用十分广泛。

用 PDF 制作的电子期刊标签，可以不受任何限制。无论多少内容、什么设计，它都能处理，而不用担心会出现掉色、污点、篇幅等问题。在传统期刊印制中，处理多色彩或非标准印品常常是很费钱费力的事。加入一个签名就可能挤掉一段文字，更不要说还有发行的问题。PDF 可以轻而易举地解决以上问题，快速完成 PageMaker、InDesign、QuarkXPress 等排版文件的转换，通过网络原样发到每个读者的眼前。

PDF 可在任何平台下开启。PDF 文件可通过多种方式得到。这里主要介绍 3 种：第一种方法是直接利用 PDF

Write 驱动程序，把任意格式的文件通过打印的方式生成 PDF 文件；第二种方法是使用 Distiller 软件把 PostScript 或 EPS 文件转化为 PDF 文件；第三种方法是使用某些应用软件的内置功能，自动生成 PDF 文件。如 CorelDRAW

8.0 版本的软件，便可通过文件保存命令自动生成一个 PDF 文件。当然包装防伪，也可利用转换软件对原有的 PDF 文件进行修改，并得到一个全新的 PDF 文件。

可扩展标记语言 XML（Extensible Markup Language）是从标准通用语言 SGML（Standard Generalized

Markup

Language）发展而来的一种新的描述型标记语言。XML 是一个精简的 SGML 语言，它去除了 SGML 中的一些非常复杂、使用率低的特性，保留了 SGML 的可扩展性能以及结构化和数据确认方面的优点，将 SGML 的丰富功能和 HTML 的易用性结合在一起，从而更适合在网络环境下使用。

XML 具备以下的特征：首先，XML 是一种定义标记语言的元标记语言。它支持用户自定义标记，通过嵌套满足一定逻辑关系的元素来组织数据。XML 的分析程序能处理所有新建标记语言。其次，XML 是面向数据而非显示的，强调内容和形式分离，使得数据显示更加灵活,同时 XML 文档集中于数据的性质与结构描述，使得相关的 XML 数据搜索更加简单高效地进行，方便信息的再利用。另外，XML 还可以标注各种文字、图像甚至二进

制文件，便于不同系统之间的信息传输，轻松实现数据的跨平台。

XML 主要由 3 个要素组成：模式（Schema）、可扩展样式语言（XSL）和可链接语言（XLL）。其中 Schema 规定了 XML 文档的逻辑结构，定义了 XML 文档中的元素、元素的属性和属性之间的关系。XSL 用来控制 XML 文档在显示时的版面风格，是一种显示 XML 文件的规范，XSL 处理器按 XSL 样式读取 XML 文件的信息。XLL 是 XML 的链接语言，支持可扩展链接和多方向链接。

XML 和 PDF 流程整合研究动机

早在 1999 年出版印刷，美国最大的图书销售公司 Barnes &

Noble 就预言，很有必要建立一套转换操作来支持他们的电子出版和按需印刷业务。目前该公司通过对 Gemstar 公司的投资，已经基本实现了基于电子出版的转换操作计划，而且他们努力和微软合作，开发出他们自己的支持微软阅读器的平台，并和 Glassbook 公司合作，在他们的网站上提供 Glassbook 阅读器的下载服务。此外，该公司正在和 IBM 公司接触，在孟菲斯的配送操作中心安装 InfoPrint

4000 和 InfoColor 70 设备，来实现按需印刷。

Barnes &

Noble 公司在全球寻找建立该操作中心的合作伙伴，但是很难找到一个价格合适、质量又令他们满意的厂家。在高质量的出版贸易和商业标牌制作中，他们要求的精确度高达 99.998%，这远远高于传统照排机提供的精度。

他们的操作只关注于大量相同内容形式的硬拷贝目录单和电子文档之间的转换。书籍出版工业只是最近才开始转向全数字化工作流程，而且对于大多数的交易清单，出版商只有在该活件全部印完时出版印刷，才能够拿到。

这种方法就是和那些生产自动标记和新格式化电子文档的高科技公司直接进行联系。随着越来越多的公司关注这一领域，我们可以预言，在短期内就有机会得到价格便宜的解决方案。一旦文本和图像硬拷贝的转换问题得到解决，人们很快就会解决其向电子文档的转换问题，要么直接进行 PS 版，要么和软件供应商联合攻关。

在对许多的选择进行评估后，Barnes &

Noble 公司发现找不到可以提供合适价钱、质量和服务的公司，于是他们自己在纽约、墨西哥城、马尼拉建立了操作分厂。

最原始的转换流程

在当前的电子出版、按需印刷和配送发行中知识产权，最终的文件传递形式都是采用 PDF 或 HTML 形式。PDF 可以要求用来推动按需印刷和阅读器设备的发展，而 HTML 在火箭电子书、软件书阅读器、标准 PC 浏览器上都有所使用。所以我们首先关注的焦点是如何完善硬拷贝和电子文档不同页面和形式的融合过程。目前，这一问题是通过对 Quark 和 HTML 流程的混合，直接产生出手工操作中所需要的变量来实现的。

为了充分利用该流程的经济性，将其变为国际性的流程实施就显得十分必要。下面举例说明。

某书店在纽约从出版商手里接到书的原稿，第一步要经过预审，主要是确认所要求的样式和一些必须更改的细节，然后会被送到墨西哥城进行扫描。

第二步，墨西哥城在扫描过程中，到底是选用 300dpi 还是 600dpi 分辨力的扫描仪，主要取决于最终所需出版形式的要求。如果出版商要求进行，则要使用 600dpi 分辨力的扫描仪；如果出版商要求电子出版的形式，则采用 300dpi 分辨力的扫描仪即可，在电子出版的工作流程中，这些扫描内容都转成 TIFF 格式的图像光盘印刷，传到马尼拉的转化操作中心。

第三步，在马尼拉的转化操作中心会完成许多很重要的工作，文件都是分区存放，图像被交到图像处理程序，文本进行光学字符识别等。然后再经过一系列的人工智能辅助系统，将光学字符识别流编辑成高质量的 RTF 格式，这样只需要很小的内存。一旦高质量的 RTF 格式转换完成后，文件就会被分成标记和页面两个部分进行接下来的操作。

第四步，在标记部分的处理中，RTF 格式都变成了 HTML 格式标签，而且还可以使用常规的 HTML 编辑工具对附加的文件标记部分进行编辑。而且此时还会产生许多新的文件，文件形式的种类取决于出版商的要求。除了产生 HTML 格式外，还可以生成 OEB、火箭电子书、软件书阅读器、标准微软浏览器等不同的格式文件。

第五步，在页面部分的处理中，文件被输入到 Quark 系统中知识产权，利用标准的页面排版技术，编排成为一个可以被纽约出版商接受的标准的 PDF 印刷样本形式。然后该 PDF 文件根据按需出版或者 PDF 电子书的需求变换成许多不同的格式。

第六步，最后文件被传回到纽约，在传到出版商手上之前，进行最后一道工序——打样。

改进后的流程

电子书行业的变化很快色彩，有很多不足的 HTML 格式慢慢就被 XML 标准替代了。到目前为止，有两种标准是通用的：大多数电子书行业中的 OEB 格式和页面展示的 PDF 格式。实际上 XML 正在慢慢地与 Frame 融合，它们的许多功能都相通。在马尼拉转化操作中心的这些转变和功能的增加，促进了从一条龙式的产生多种输出样式的手工出版流程到采用两步式的产生同样多样式的自动出版流程的转变。

由于有了 RTF 格式，自动流程和手工流程有很大程度的相似性。最大的区别在于 RTF 后处理过程。标记部分被处理成为标准的 DTD 格式，而且被储存在 XML 库中，而代替了传统的 HTML 格式。由于没有证据证实 XSL 和其他 XML 能够产生符合印刷质量要求产品的能力，所有实际生产中文件数据被毫无选择地储存为两种不同形式的文档：XML 版本用来产生不同形式的非页面输出，PDF 版本用来产生页面输出。而且在使用 XSL 的过程中，像 OEB、HTML4.0 格式的输出也有可能被用到。

存在的问题

在该流程的使用过程中会碰到一些问题，这些问题也许 XML 的从业者都碰到过。包括从使用 XSL 创建高质量的页面和管理大量的 XSL 风格转变类型，到围绕在过程中应该标记的地方，在马尼拉转化操作中心缺少产品检测工具的环境，甚至没有一个可靠的 XML 运行基础可以提供等问题。

1.创建高质量的页面

创建高质量的页面是第一个大的挑战。出版业长期致力于通过应用高质量组合的自动化而减少或消除设计方面的干扰。这方面和 XML 流程基本上没有什么差别。而内在的问题是，XML 对文本是非页面展示的，如果对页面产品有强烈的需求，很多复杂的问题都将留到下一步去解决。于是就会产生更多的具有挑战性的问题：

（1）跨页面是隐藏的孤行、字控制；

（2）页面内和跨页面的行平衡；

（3）字距调整、字间空格调整；

（4）无文本元素和文本元素的并列布置；

（5）复制过程中对原稿的保存等。

由于使用了 XSL 格式代码将库中标记的 XML 转移到最终的目标格式，在该环境下我们的操作系统可以达到最大适应性。但是对于用 XSL 取代从 XML 到高质量页面的

全过程，人们都还持怀疑的态度。

2.XSL 增殖

XSL 增殖是另一个必须考虑的问题。为了得到各种形式的输出，我们往往要使用一种格式的 XSL，而且很有必要使用更多的、更成熟的格式进行更多样式的输出，满足更多出版商的特殊要求。这就要求样式表单具有两种独立的功能。一是转换功能，如 XML 转换成为 HTML 格式等；二是花式功能，即使得多种不同形式的输出看起来与出版商要求的样式一样，或者自己与样书进行匹配。

3.迭代标记

在上面所描述的过程中标签，有多处用到了迭代标记：最开始的分区中，在对 RTF 的处理过程中，在相对较纯净的 XML 环境中进行后处理过程中。这里就存在着对生产效率和精度的选择问题，最好的方法是尽可能接触文本和图像，而且要从接触过程中获得尽可能多的信息。举个例子说纸箱纸盒，图像在扫描之后就必须对页面图像进行分区，并且进行光学字符识别和图像处理。在文本元素可以识别的基础上，通过使那些区域和文本元素关联起来，进行标记的保存。完成这项工作并不需要在整个过程中做很多额外的工作，而且操作者也很难提供有问题的、与后面使用的 DTD 格式不兼容的标记。

4.产品控制工具

Barnes &

Noble 公司在马尼拉的操作中心严格按照标准、过程控制等创建了一个高效率的生产环境。在这个工作环境中，操作者可以使用工具严格地定制工作步骤，简化工作程序，达到高效和高质量的要求。更进一步的是，这些工具可以整合进整个工作流程中，而且还可以推动整个过程的产品控制和内容数据库管理工作。

结束语

在过去几年的实践中重组，Barnes & Noble 公司越来越意识到建立一个世界级的转换操作中心的必要性，用来进行页面和标记内容的整合非常重要。转换操作中心也正在运行中，人们也在慢慢地适应它，逐步解决围绕在 XML 和 PDF 流程整合过程中产生的各种问题，并走向成熟。通过努力发展史，Barnes & Noble 公司感觉到他们帮助整个行业实现了许多人正在分享的目标：任何一个客户无论在何时何地都可以方便地找到适合自己使用的阅读形式的书籍；出版商也有希望去生产任何他们中意的电子书籍样式，而不再受到对原稿内容进行再改造的局限。

在电子出版和按需印刷在出版业所占的比重越来越大的时代，相信利用科技解决数码印刷中各流程的整合问题，充分利用设备，提高效率意义重大现状及趋势，也势在必行。