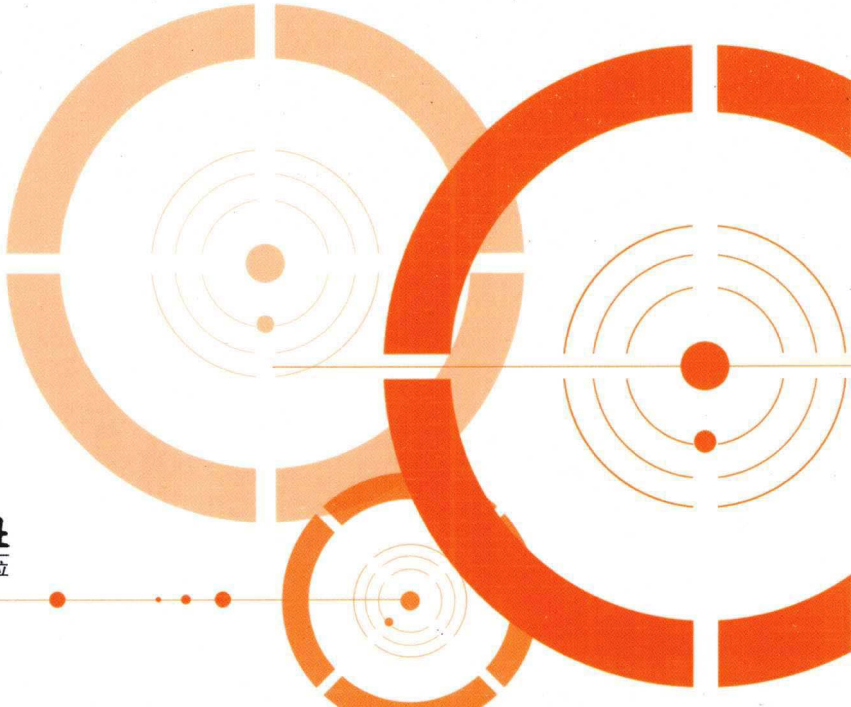
数字出版系列丛书

数字出版实用教程

SHUZICHUBAN
SHIYONG JIAOCHENG

黄孝章 张志林 陈功明◎著



知识产权出版社
全国百佳图书出版单位

数字出版系列丛书

数字出版实用教程

黄孝章　张志林　陈功明　著

知识产权出版社
全国百佳图书出版单位

内容提要

本教程编著主旨和架构原则是重在实际应用，基于先进的动态数字出版平台系统，重点对从数字化加工、生产管理到数字化运营等全流程实际进行操作性讲解，帮助出版单位领导人员了解数字出版平台体系的总体架构和具体模块功能；帮助相关院校专业学生和数字出版具体岗位人员掌握数字出版的实际运作。本书可以作为学校相关专业教学实训教材，也可以作为行业数字出版岗位人员培训教材，以及实际操作指南。

责任编辑：于晓菲　李德升

图书在版编目（CIP）数据

数字出版实用教程/黄孝章，张志林，陈功明著. —北京：知识产权出版社，2013.2

ISBN 978-7-5130-1706-0

Ⅰ.①数… Ⅱ.①张…②黄…③陈… Ⅲ.①电子出版物—出版工作—指南 Ⅳ.①G237.6-62

中国版本图书馆 CIP 数据核字（2012）第 268444 号

数字出版系列丛书

数字出版实用教程

SHUZI CHUBAN SHIYONG JIAOCHENG

黄孝章　张志林　陈功明　著

出版发行：知识产权出版社

社　　址：北京市海淀区马甸南村 1 号　　**邮　　编：**100088

网　　址：http://www.ipph.cn　　**邮　　箱：**rqyuxiaofei@163.com

发行电话：010-82000893 转 8101　　**传　　真：**010-82005070/82000893

责编电话：010-82000860 转 8363　　**责编邮箱：**yuxiaofei@cnipr.com

印　　刷：知识产权出版社电子制印中心　　**经　　销：**新华书店及相关销售网点

开　　本：787mm×1092mm　1/16　　**印　　张：**13

版　　次：2013 年 2 月第 1 版　　**印　　次：**2013 年 2 月第 1 次印刷

字　　数：205 千字　　**定　　价：**38.00 元

ISBN 978-7-5130-1706-0/G·536（4548）

前　言

目前来看，开展数字出版还缺乏基础性、共性技术的实际运作平台的借鉴，缺乏能深入到实际操作层面的内容处理技术的借鉴。本书主要针对出版社数字化转型的需要，针对现阶段出版社，尤其是图书出版社（出版集团）开展数字出版的应用需求，结合目前先进的数字出版内容管理平台进行实际运用操作性讲授，希望对出版社开展数字出版业务及从事内容资源建设的新媒体企业有所裨益。

本书编排面向出版产业链的变革，贯穿出版社从内容采集、内容创作、内容聚合以及内容传播等生产经营的全环节，尤其是从数字出版内容平台的内容聚合与内容分发角度组织全篇。在出版运营环节，出版社可以借鉴网络运营已经成熟和正在不断产生的商业模式，这方面有各行业共性的理论和经验可资借鉴，本书也会观察总结业界的新鲜经验。在出版内容的生产环节，需要对内容资源的结构化加工、生产进行详解，并引导实际操作。

本书的定位不在理论探讨，而在数字出版的实际应用。以出版社历史资源数字化加工的需求形态为起点，重点讲解数字化内容的生产与内容资源的管理，通过基于 XML 结构的信息内容处理技术的详细讲解，为知识的重用以及再生服务打好基础。最终实现出版社内容资源适应网络新媒体特点的传播，为出版社带来效益提升的目标。

本书编排特点是，将出版社需求、通用数字出版技术与国家出版重点工程项目有关课题相对应，从实际运作出发，每一章提出基本问题，界定相关概念，以图示引导讲解功能实现方式，方便实际操作。最后，对基于本教程讲述的数字出版系统尚未涉及的新应用进行简介，以拓展数字出版技术应用的范围。

本书内容部分为北京印刷学院重点项目（“三网融合”背景下北京数字内容产业发展模式研究）及北京出版产业文化研究基地项目（北京数字出版产业商业模式及平台建构策略研究）的研究成果，同时得到了上述两项目的出版资助。

由于时间仓促，书中观点和内容难免有错误和不妥之处，敬请读者批评指正。

张志林

2012 年 11 月 20 日于北京印刷学院

目　录

第 1 章　数字化加工

1.1　信息技术应用的简要回顾

王选教授曾经指出，印刷出版行业是最早使用计算机的行业之一，但是这个行业的信息化发展却比较缓慢。究其原因，其他部门的信息化建设，在产品层面只是其产品元数据的数字化，产品本身的物理性质并没有改变；而出版业的信息化是内容生产的信息化，不仅产品元数据数字化，而且产品本身也数字化，从原子形态变成比特形态了。因此，中文信息处理技术是出版行业最重要的信息化技术。

1.1.1　中文信息处理技术的应用

虽然出版行业应用计算机进行排版已有几十年历史了，但印刷出版和电子出版之间，很长时间里软件技术互不兼容，绝大部分常用的文字编辑排版软件都是面向打印和印刷的。印刷排版软件技术的着眼点在于图文呈现，关注字体清晰，印刷精美，印刷版的“人性化因素”使图书不需要任何特殊设备便可阅读，便于注释。从数字技术的角度考察，印刷版的排版软件处理对象为线性结构文件，不具备检索、重用的功能。其后的电子出版发展，产生了全文数据库电子出版物。数据库软件的处理对象是数字化的结构性文件，它能够提供数据存储和分析能力。电子出版的优势是可以进行数据的索引、排序、查找、在线浏览，相关一致性检查，可以提供强大的检索功能。

尽管两种出版方式都使用计算机系统，但是两种文本的显示格式不能互通，这就提出了如何使创建的数字内容既可以用于印刷版，又可以用于电子版

的数字技术要求。解决这个问题的途径是将这些软件输出的结果数据进行归一化处理，将这些数据转换成为 XML 文件格式❶存储到内容资源管理系统（Content Resource Management System，CRM）中，经过出版引擎实现出版产品的跨媒体发布。出版信息的 XML 结构化处理、XML 数据的存储与检索以及出版信息的跨媒体发布，构成了该系统的技术核心。如图 1－1 所示。

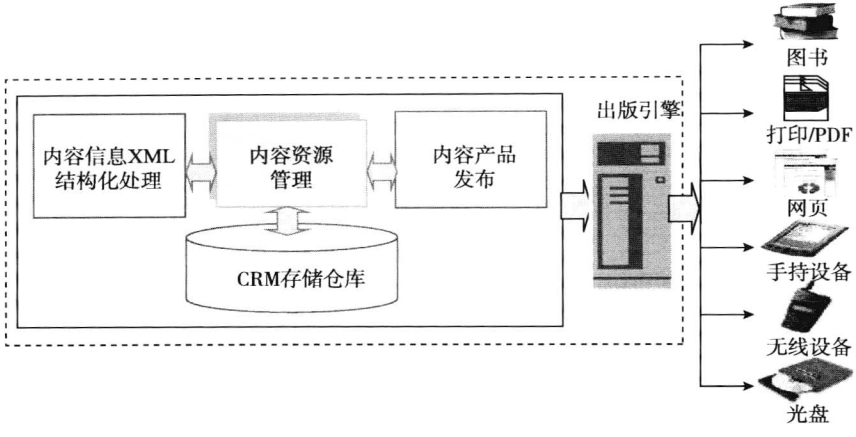


图 1－1　内容信息归一化处理跨媒体发布

CRM 系统适合跨媒体出版的要求。通过 CRM 系统，出版信息的跨媒体检索、重用、挖掘、交易能够展开，使出版信息的生命周期得到有效的延长。因此，实现 CRM 系统中文档格式转换的自动处理，一直是人们追求的目标。但目前的技术离人们的期望还有相当的距离，以至于大量的数据转换需求，衍生出了一个专门从事转换数据格式的数据加工行业。

CRM 技术与跨媒体出版技术的有机结合，给数字出版技术的发展描绘了巨大的发展空间。数字出版技术属于信息技术的研究与应用领域，因而数字出版的本质规定性是对内容的深度加工、分类与整合，是内容的信息化而不是简单的电子化和屏幕化。中文信息智能化处理是计算机中文信息处理的基础性研

❶ 可扩展标记语言（Extensible Markup Language，XML），用于标记电子文件使其具有结构性的标记语言，可以用来标记数据、定义数据类型，是一种允许用户对自己的标记语言进行定义的源语言。XML 是标准通用标记语言（SGML）的子集，非常适合 Web 传输。XML 提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据。

究，是信息处理的关键技术。有许多研究人员在这一领域耕耘攻关，但至今仍处在技术突破的前夜。由于中文信息处理的特殊性，解决中文信息的复合出版、自动标引、自动分类、自动聚合、智能检索，只有在中文出版的过程中才能实现；只有解决了数字出版关键技术，才能最终解决出版行业信息化落后的状况。

1.1.2　XML 成为跨媒体出版的重要标准

20 世纪 80 年代后期，全文数据库出版解决了电子版格式与打印版格式的统一问题，即检索结果显示格式与排版印刷格式的一致性。解决的方法是，每篇文档每一页同时有两种版本：扫描图像版本和用 SGML 标记语言的文字版本[1]。标记是隐形的，最终在网页浏览或者阅读器上浏览，看不到标记本身，而只有标记的结果。扫描图像使用与印刷版相同的图案和布局，能够保证每页的打印和显示与印刷版一致。由于扫描图像没有电子版所需要的属性标识，文档中的各种排版符号依然保留，因而不能直接在计算机上浏览。采用 SGML 标记语言将文档的内容与样式分开，正文则用于建立全文索引，以便信息检索和在屏幕上的快速显示。

在美国，早期的 SGML 使用者是科技期刊出版商，由于科技期刊内容存在大量复杂的公式、表格和图片，版面复杂，无论是印刷版还是电子版，都是最难制作的出版物。学术期刊比其他出版领域更多的使用 SGML，芝加哥大学出版社的《天体物理学期刊》从 1994 年开始采用基于 SGML 工作流程，1995 年开展网络出版时，所有的 SGML 文件转换成 HTML 文件[2]，在很短时间里几乎没有增加额外成本就在互联网上出版这本大型复杂的期刊，实现了双轨出版。

---

[1] SGML（Standard Generalized Markup Language），即标准通用标记语言，于 1986 年出版发布的信息管理方面的国际标准（ISO 8879）。该标准定义独立于平台和应用的文本文档的格式、索引和链接信息，为用户提供一种类似于语法的机制，用来定义文档的结构和指示文档结构的标签。其中 Markup 的含义是指插入到文档中的标记。制定 SGML 的基本思想是把文档的内容与样式分开。SGML 是一种在 Web 发明之前就已存在的用标记来描述文档资料的通用语言。但 SGML 十分庞大且难于学习和使用。

[2] HTML（Hypertext Markup Language），即超文本标记语言，是用于描述网页文档的一种标记语言。互联网是建立在超文本基础之上的，文本中包含了“超级链接”点。所谓超级链接，就是一种 URL 指针，通过激活（点击）它，可使浏览器方便地获取新的网页。这也是 HTML 获得广泛应用的最重要原因之一。

现在，学术期刊以及其他出版则采用 SGML 的后续技术 XML（Extensible Markup Language），即可扩展标记语言。

出版信息的 XML 结构化，为实现自动的跨媒体出版打下良好的数据基础。XML 数据配上用于显示 XSLT 的样式数据❶，解决了出版信息在互联网上的发布。但是，XML 数据解决出版信息到印刷和光盘的发布还存在障碍，特别是在 RIP❷ 不支持 XML 输出的情况下，出版信息到印刷和光盘的跨媒体还难以实现。解决这个瓶颈的一种方法是，通过软插件技术将 XML 数据直接嵌入到排版软件的版面上，由排版软件实现 XML 数据到 PS（Photoshop，像素制图软件）❸ 数据的转移。纸介质的出版信息发布问题解决后，将 PS 数据转换为 PDF 格式❹，光盘发布的问题也迎刃而解❺。

21 世纪开启，在出版领域，为一种出版物同时制作电子版与印刷版的文档处理软件使跨媒体出版开始流行。运用 XML 标记语言规定的元数据结构，实现了新闻信息的内容描述、交换和再利用，电子文件能按照不同的方式呈现内容，既可在屏幕上显示，也可用于印刷。在报业，出版纸质报纸、网页新闻以及 2005 年以后广泛流行的手机报，形成“纸网互动、滚动报道”的立体报群传播态势，技术上都要归功于 XML 成为中文新闻信息置标语言在报业

---

❶ XSLT（Extensible Style sheet Language Transformations）。在计算机科学中是扩展样式表转换语言的简称，这是一种对 XML 文档进行转化的语言，XSLT 中的 T 代表英语中的“转换”（transformation），它是可扩展样式表语言 XSL 规范的一部分。

❷ RIP（Raster Image Processor）光栅图像处理器，对于计算机直接制版系统来说，RIP 的主要作用是将计算机制作版面中的各种图像、图形和文字，解释成打印机或照排机能够记录的点阵信息，然后控制打印机或照排机将图像点阵信息记录在纸上或胶片上。

❸ PS（Photoshop，像素制图软件），是 Adobe 公司旗下最出名的图像处理软件之一。它不仅是一个很好的图像编辑软件，同时它的应用领域很广泛，在图像、图形、文字、视频、出版各方面都有涉及。

❹ PDF（Portable Document Format）是 Adobe 公司开发的电子文件格式。该格式与操作系统平台无关，这一特点使它成为在 Internet 上进行电子文档发行和数字化信息传播的理想文档格式。PDF 文件格式可以将文字、字型、格式、颜色及独立于设备和分辨率的图形图像等封装在一个文件中。该格式文件还可以包含超文本链接、声音和动态影像等电子信息，支持特长文件，集成度和安全可靠性都较高。经中国国家标准化管理委员会批准，PDF 文件格式已成为正式的中国国家标准，并已于 2009 年 9 月 1 日起正式实施。

❺ 肖建国．XML 和 DAM 技术与跨媒体出版．中国印刷．2001（4）6－7

的应用[1]。

目前，一些中文处理软件能够直接利用排版软件产生电子文本，加工成计算机可读的电子书，其加工过程不是在排版文本产生之后而是融合在排版过程之中。也有出版社开发出利用数字化加工的 PDF 文件，自动生成图书的 XML 元数据信息的转换软件，在跨媒体出版方面有了自主开发的信息处理技术。

1.2　什么是数字化加工

传统出版企业认识到，文化产业的发展核心是内容，向数字化转型是出版企业发展的必由之路。实现传统出版与数字出版的融合发展，需要利用各种类型的信息加工方式，完成存量出版资源的数字化整理加工，加快数字内容资源的深度整合，加速内容的数字化。在此基础上，实现内容资源全方位、深层次的开发利用，并借助互联网、手机、电子书阅读器、平板电脑等新型传播途径，重构知识与内容的销售渠道。对数字化加工的理解，表现了业界对数字出版本质认识的不断深化过程。

1.2.1　数字化加工的内容

出版资源的数字化加工，是指对出版信息资源的数字化整理，主要是完成对传统资源的加工、分类和标引工作。数字化加工包括了两部分内容，一是对已经形成纸质图书的历史出版资源重新进行电子化、代码化识别、审校、重排、标引；二是对目前已经电子化、代码化的内容进行基础标引和各种基于专业需求的深度标引。

进行数字资源加工，首先要对所采集的内容进行数字化转换（OCR/SCAN）处理[2]，然后进行人工标引加工处理。现在，有的出版社已经自主开

---

[1] 新闻领域的中文信息处理率先实现了 XML 标引，并已经颁发了国家标准：GB/T20092—2006 中文新闻信息置标语言（Chinese news markup language，CNML）。

[2] OCR（Optical Character Recognition），即光学字符识别；SCAN，即扫描。通过光学扫描仪和计算机的配合，OCR 软件将图像数据进行运算分类后转化为计算机内码。它可以极大地减轻数据录入工作的强度，提高数据录入的速度。

发出多核心 OCR 数据加工生产线，具有完备的流水线式操作体系和管理监控系统。更进一步，信息处理技术能够自动提取其中的标注（Tagged）内容及全文文本，对内容进行过滤、分类或自动摘要。最终，经过标引的内容转化成内部标准格式，并与其关联的对象（包括图片、原版式等）一起装入内容仓库中存储，供查询、挖掘及其应用。

1.2.2 数字化加工的作用

对出版社开展数字出版来说，将出版社的存量资源和新产出的增量资源内容进行电子化、代码化入库，进行基于整书元数据的 XML 标引，它的作用在于建设 CRM 系统，实现内容信息的跨媒体出版。历史资源的数字化整理加工，是一件最基础、最基本的工作，每种出版物的全部信息都要通过数字化加工进入 CRM 系统，实现从一个入口直接找到相关的信息；对系列丛书能够有效关联，甚至实现资源之间的有效关联；对多版本信息也要各自独立加工、入库管理，并且不同版本之间也能够进行有效关联。

数字化信息资源的建设与管理对现有印刷品的数字化需求越来越强，OCR 技术应用成为 CRM 系统建设中的重要阶段，同时也是数据加工的核心技术。经 OCR 技术处理的电子文档，可广泛应用于各种电子出版物、网络资源、各种大型文献资料数据库、数字图书馆等众多领域，也是出版社内容信息资源开发利用的必经阶段。

由于出版社对存量出版资源的数字化整理加工的需求不一样，因此，加工的层次有初级和高级之分。最常见经过数字化加工的图像 PDF 文件格式，能够将纸质文档转换成图像文档进行阅读；或者进一步，将文档内容转换成计算机代码，保持图书的原版原式进行阅读。通常，这种实现了纸质文档向图像文档转换的图书内容，能够满足出版社最基本的出版资源电子化、代码化需求，以及读者最基础的阅读需求。为了满足未来数字出版新商业模式对数字出版产品的要求，还需要实现元数据内容的自定义和可扩展，以及基于内容的深度标引，以期能够满足“一次制作、多个渠道、重复使用”的跨媒体、跨渠道出版需要。

通过数字化加工，使内容能够在纸本上、屏幕上显示阅读，使之不断开