我国科技期刊应尽快实现基于结构化排版的生产流程再造*

刘 冰 游苏宁
中华医学会杂志社,100710,北京

摘 要 从我国科技期刊数据生产当前的阶段性特征和产业发展的要求入手,以企业流程再造理论为依据,围绕如何提高内容生产和数字资产管理及信息服务效率,就我国目前科技期刊数字化发展中的数据加工和内容管理的运行模式,结合当前的国内外实践进展,提出了基于可扩展标志语言的结构化排版的科技期刊数字化生产流程模式概要构想和实施等问题。

关键词 科技期刊;流程再造;结构化排版

Chinese sci-tech periodicals should be as quickly as possible reengineering publishing process based on structured typesetting with XML//LIU Bing, YOU Suning

Abstract Based on the theory of process reengineering, the study describes current characteristics of China's sci-tech periodicals typesetting. It also describes requirements of database information publishing and concepts of structural publishing. XML can organize the information and represent the knowledge in a more structural manner. Furthermore, it can present the documents in multiple layouts with the pre- formatted templates or style-sheets. Chinese sci-tech periodicals should be as quickly as possible reengineering publishing process based on structured typesetting with the XML, to improve operation efficiency of simultaneous cross-media and multichannel publishing process.

Key words sci-tech periodical; process reengineering; structured typesetting with the XML

Authors' address Publishing House of Chinese Medical Association, 100710, Beijing, China

20 世纪 90 年代, Michael Hammer 和 James Champy 提出的流程再造理论认为,对企业的业务流程进行根本性的再思考和彻底性的再设计,企业可以在成本、质量、服务和速度等方面取得显著的改善[1]。迈入数字时代,传统科技期刊出版在转型过程中面临的最大挑战是如何有效率地将各类型资源进行生产、整理、交换、推广及提供优质的信息服务。要适应市场环境和技术手段的进步,就必须对传统的生产流程进行思考与再设计,提高期刊的生产和信息发布效率,增强核心竞争力,实现新的战略发展目标。这些有赖于有效的标准化的技术来解决。当前,数字出版领域没有统一的行业标准,各大著名出版社基本上都是采用基于可扩展标志语言(eXtensible Markup Language, XML)技术对数据进行结构化生产,为目前和将来的动态或复合出版工作流程奠定运行和发展的基础。

1 结构化排版概况

1.1 什么是结构化排版 笔者以为,结构化排版是指借助计算机软件和语言技术,通过建立规范的信息格式和标准,将期刊内容与样式分离进行实时的结构化生产和编辑,使得经过标引的文件或元数据形成资源储备,在独立于文件格式之下实时变更、适应不同媒介进行发布的一套现代出版生产体系。它具有以下特点:1)一次制作,完成元数据的深度、细化的加工和标引,有效提高信息处理和检索速度,强化内部创作协同一体化;2)多元发布,格式多样化地适应不同展现层的信息传播媒介或终端设备,实现跨媒介出版;3)重复可用,有效挖掘、提炼资源,加强数字资源整合;4)规范标准,统一出版标准的数据保证信息源一致、准确、规范,利于进行各种格式的转换、传输、存储,提升出版效率。

1.2 基于 XML 的结构化排版能够实现内容、结构与展现的完美统一 万维网联盟制定的 XML 标准是一种简单的数据存储语言,是继超文本置标语言(HTML)之后新一代网络整合技术,是建立结构化文件和数据的通用格式。其文档是一个文本文件,使用一系列简单的、可用方便的方式建立的标记描述数据,易于掌握和使用。HTML 是当前电子文件的标准规格,着重于版面编排与外观格式,对于文件结构的规范及内容语意的描述乏善可陈。XML 弥补了 HTML 可扩展性和结构性的不足,加之其开放性和灵活性,为数字出版带来了革命性变革。XML 可作为各种媒介间转换的桥梁,在数字出版、电子商务、数字博物馆或图书馆、电子数据交换等领域展现出强大的应用潜能。XML 可按照统一的格式灵活采集各类电子信息到内容资源库中进行管理、加工、整合和共享,在各种媒介进行内容发布或按需印刷[2]。

1.3 国际领先出版集团科技期刊出版的结构化排版情况 国际领先的出版集团的期刊数字出版平台集成了学术期刊数字资源采集、加工、存储、后台管理、信息

* 中国科协精品科技期刊工程项目(QK2009003-A)

服务等多项功能。

Springer公司的SpringerLink拥有52家出版机构的1 900余种期刊,其资源加工通过Aries公司的“在线同行评议系统”和“编辑管理系统”实现,以此为核心节点,实现内容的顺畅生产和流转。同行评议后的稿件进入数字化编辑加工流程,经标记和结构化的内容可直接在线出版和形成用于印刷的高质量PDF文件。还可进一步挖掘、组织和定制数字化服务产品,通过数字化决算支付系统实现利润[3]。

PubMed Central(PMC)是美国国家医学图书馆(NLM)的国家生物技术信息中心建立的生命科学期刊全文数据库。当前,被PMC收录的期刊在上传电子数据时要求转换为XML格式[4]。Medline数据库是NLM旗下MEDLARS系统30多个数据库中最大的著名的生物医学数据库,被其收录的期刊题录和摘要数据的上传也需按XML格式制作标签数据(对文章作者、题名、出版日期等各种信息附上专门定义的标签或定义符),进行电子提交,通过审核后才能收录进入MEDLINE/PubMed检索系统。对于提交XML电子数据的杂志社,NLM还提供与杂志社网站期刊论文全文的链接[5]。与传统的键盘录入或扫描相比,期刊发布和传播效率大幅度提高。

Charlesworth集团中国公司为期刊提供以XML为基础的工作流程服务,工作内容包括数据处理、排版、制图、版权服务等,为50多家出版商(如Nature、BMJ、Blackwell、Taylor & Francis等)生产加工近400种期刊,排版以XML形式完成,提前出版网络版,现阶段BMJ文章自动排版率超过93%,提高了出版效率,降低了生产成本。同时可快速进行全球有效而稳定的传输[6]。

1.4 结构化排版的流程和主要技术要求 结构化排版需在生产初期提供写作模板初步规范数据,后期进行数据转换以便进一步生产加工。

1)模板。国际大型出版商的多数在线投稿和审稿系统为作者提供了成熟的写作模板,通过Word模板或基于科技排版软件Latex实现[7]。

2)从Word到XML的转换。在线编审平台提交符合出版条件的稿件按照数据格式标准通过Word to XML的转换实现结构化排版,最终多元、多次发布。

3)基于XML的排版。专用的排版工具LaTex、Adobe Indesign已使用得比较成熟,如Elsevier、Springer、Cambridge University Press、Kluwer Academic Publishers等出版机构均接受LaTex排版的稿件[6],利用完整的XML及HTML输出功能,为科技期刊出版数字化出版策略助力。

4)技术商在不断的提升公式、化学式、表格等特殊内容元素的解决方案,建立与XML密切结合的数学、化学等特殊学科的输入专用工具。

2 结构化排版的价值与意义

XML结构化排版文本因其灵活的可定制的数据结构可以使出版的内容具有高效的互操作性、可访问性和重用性;1次设计多次重用的样式可以使内容与样式分开,轻松发布在各种媒体上;整合及关联的内容可以提高内容资源平台的灵活性,实现资源的多次利用和快速增值。

2.1 提升出版效率 成熟的XML编辑器,可以与已有的在线同行评审系统整合使用,能更快提供完整的文档;从开始使用XML编辑拥有对页面安排的控制权,准确跟踪、自动执行文档工作流程,所见即所得方式实现内容撰写、内容协同,使版式更加美观统一。系统也可以方便地为编辑和作者建立个性化的标引体系。排版后的内容可以直接对应传统印刷、按需印刷,加速生产进程。

2.2 实现数字资产管理 排版过程和结果保存XML文件、样式文件以及视频、图片文件,XML文件和样式文件能够很方便地转换成PDF、HTML、手机格式等各种内容格式。XML文件的高度结构化和有效表达,可以很容易与目前已发展成熟的各种数据库管理系统进行数据交换和整合,结合相应的处理图片、图像、音视频工具,可对多媒体资源进行专业化管理。

成熟的XML工具还可以实现网络资源采集以及进行符合数据标准的转换。纸质出版形成的海量历史资源通过合适的XML标引工具进行标引,自动进行格式转换加工,形成可重用的颗粒化、碎片化内容。这些数据进入资源管理库,通过统一和有效管理实现强大功能,如版权重用、版权保护、版本控制、在线服务等,促使科技期刊出版从信息处理阶段跨越到知识管理阶段。

2.3 便于数据的多元和多次发布 新媒介和阅读终端的涌现极大地丰富了读者获取信息、沟通和交流的渠道。新技术环境下成长起来的科技工作者几乎习惯并依赖于网络和新媒介,这给传统科技期刊的发展带来了挑战,科技期刊可持续发展的出路在于利用现有资源和技术整合与融合,不断创新。以先进的生产手段——结构化文档为未来的内容服务提供必备的基础,整合形成具有新价值和影响力的媒介,打造集纸版、网络、手机、电信等为一体的复合型传媒产业集团。国际众多科学、技术和医学出版商已经采用“在线优先”的工作流程,而XML的高质量输出和稳定性能够保证在线优先的出版策略。

2.4　提升产品竞争力　通过对章节内容、文献、图表、公式定理等自动编号,建立参考文献和引用库等方式,XML 文件保存了内容的结构和相关信息,方便未来对内容的再利用和挖掘,例如抽提题录信息、年度关键词和目次索引、生成知识库等。单就信息检索来说,XML 具有自我描述性质,可以提供语意层次的搜寻,提高检索结果的精确率,输出各种类型需求的内容,目次、参考文献、公式、定理、插图、视频、页码等各类基础信息可实现任意形式的交叉引用。有效的数据管理,可以保证内容的完全一致,组织并发布不同形式的内容,从读者的实际需要出发实现出版资源的多样化,由专业出版向专业大媒体转变,提升产品的竞争力。

2.5　为信息服务系统开发提供储备　随着信息服务渠道和手段的增加,借助 XML 的可扩展性、数据与样式分离等特色,对储备的信息资源可以进行加值处理,通过添加各种宏包以扩大其功能,实现各种特殊要求,使得信息服务的应用和发展更具空间。标准的内容标注可以实现对内容特征的描述,为检索各种内容资源提供基础,使内容价值最大化,为正确和有效地重用内容资源,建立知识网络,实现内容价值最大化提供战略储备。国外的实践证明,传统出版向现代出版转型的过程中,数字出版具有无可比拟的高时效、高智能化的特点,科技期刊内容生产过程形成统一的数据源,可避免重复劳动,实现高频度重用。再辅之以格式化、标准化的信息格式,能保证数据准确,进而进行信息挖掘和组合实现多媒介、多形式的信息服务。

3　结构化排版的成本分析

彭玲[8]就高等教育出版社的每面的数字化生产与传统纸质生产成本的比较显示,前者是后者的 1.0 ~1.5 倍。诚然,出版社进行软硬件的投入以及后续的维护和升级,其成本必然很高,这些是制约集约化程度小的期刊社进行数字化流程再造的瓶颈;然而,考虑纸质出版、后期数字化加工、其他载体形式的出版以及集成和重组信息的发展前景,拉长生产时间会摊薄一次性投入的成本。革命性技术变革的投入是必需的,基于 XML 的结构化排版对期刊内容的数字化、元数据信息的提取、多种载体的发布等可以一步实现,与传统纸质出版后再行数字化的流程有着本质的区别。建立了完好的盈利模式后,新的出版平台和形式会带来更多的商机。

4　实现结构化排版需要做的工作

以结构化排版为基础的信息服务大有可为。通过自动组合多种信息以各种形式发布,消除了传统发布流程的低价值和劳动密集型的过程。期刊可以把“最有技术含量”(结构化排版软件的开发及信息服务系统建设)和“最没有技术含量”(数据生产及排版校对)的工作进行外包,把科技期刊产业培育形成真正的微笑曲线(施振荣的微笑曲线理论认为在产业链中,附加值更多体现在设计和销售两端,处于中间环节的制造附加值最低),加大内容策划、市场推广力度。需要做的基础工作如下。

4.1　建设文档结构定义(Document Type Defination,DTD)标准及相关模板　在 DTD 中要定义 XML 文件中所需要的信息结构和结构之间的关系。如对应一期学术期刊,〈Journal〉设置为最顶级的信息结构,〈Article〉中包含题名、作者单位、摘要、关键词、全文、参考文献等必要结构,其中题名、摘要、关键词的信息片段只能存在 1 个实体,全文段落可以存在多个。通过严格定义,XML 文件配合 DTD 构成具有特定应用特色的信息载体,允许读者进行信息检索和获得个性化信息服务。高等教育出版社定义了一套适用于学术期刊内容结构化的 DTD——HEP Journal. dtd。HEP Journal. dtd 可以描述文章、期、卷的所有元数据信息和文章内容,涵盖了人文社会科学及理工类学科,通用性较好,与国外一些期刊发布平台及国内清华同方期刊数据库等系统能无缝交换元数据信息;然而,中国科技期刊还没有在科学、技术、医学主流领域形成各自的 DTD 标准促进期刊的规模化发展[8]。此外,需要对期刊的版式进行设计,建立集群化期刊通用的模板或者个性化期刊特色模板以供选用。

4.2　现有期刊集团自建或者购买软件,加强硬件建设　实力较强的期刊社可以自主开发或外购 Word 转换为 XML 的工具和 XML 排版软件,建立 Word 文档清洗队伍。在对 XML 文档的结构和内容进行分析的基础之上,进行脱机规则学习或联机的文档清洗[9]。基于语义对文档进行合并和分割,组建面向网络服务的信息管理和服务系统,提高信息存储和检索效率。有关学者和国家层面提出了相关建设构想或课题攻关。邹振亚[10]设计了一套基于 XML 和工作流的出版发行系统,对关键技术进行了分析研究。新闻出版总署的“数字复合出版工程”提出了开发建设数字复合出版系统软件的目标,对出版业原有的只能录入排版、生产印刷版文献的计算机排版软件进行数字化改造,集成可以同时生产印刷版与数字版文献的系统软件。

内容结构化采集、编辑是出版流程的起点,是期刊产业向全媒体发展的基础,对整个出版流程能高效、准确地完成有着重要的作用。单本期刊可以通过加盟现有期刊集团的优质平台或者以外包形式委托加工内

容,加之人工标引,形成编辑、制作、出版、发行的完整产业链,推进我国出版业的信息化进程,使出版者既是出版内容的生产者,也是数字内容发布与传播的运营者,成为数字出版的主体[11]。中华医学会杂志社在各项业务系统建设完成以后,拟推动结构化排版工作,建立医学文档结构标准及样式文件,形成资源储备和成本优势。中华医学会系列杂志具有一定的规模性,出版体例整体规范统一,绝大部分无特别复杂的数学公式、化学式和复杂表格,通过现有的学术影响力和大量培训能够有效解决模板的推广问题。通过构建基于XML的数字出版核心的内容资源生产环节,中华医学会系列杂志可实现内容资源整合及增值加工,为中华医学会杂志社由传统出版商向专业的医学科技信息提供商的跨越打下坚实的基础。

2008年3月,HighWire发布的利用XML语言重建的新电子出版平台H20以最大的灵活性给使用者在各个层面注入了新的标准,能与其他系统相互关联,延伸网络服务和技术,其内容将以web2.0模式工作。其标准化的内容,已经可呈现在许多新型设备(如苹果公司的3.5英寸宽屏手机iphone)中[12]。

4.3 作者和编辑的观念培植 在传统出版模式的条件下,作者通常不能完全专注于创建和改进内容,浪费大量时间进行文档的格式化处理。通过与已有的在线同行投稿和审稿系统有效接驳,可以给作者提供写作模板。高等教育出版社在实施内容管理平台的过程中,特别注意对相关人员的培训,在加强技术力量的同时,进行了一系列的人员培训工作,如普及XML语言基础知识,培训新的XML排版人员和内容结构化人员,发展排版外包商等。人员培训和储备为从传统出版向跨媒体出版转换奠定了人才基础,有利于持续性、渐进性地改进现有的出版流程。结构化排版与现有出版流程整合后,可以在XML排版文件的基础上进行方正PS文件印刷,也可实现基于XML的印刷[13]。

4.4 与展现平台的接驳和功能建设 内容资源生命周期的最后一个环节即提供信息服务。内容展现平台可以提供电子商务,用来提供产品和服务的个性化消费支撑环境,实现内容的商务价值。

厉衍飞等[14]综合对国内外学术期刊数字出版平台的研究,提出当前国内应加强以下几方面的功能建设。

1)建设并增强知识搜索功能。

2)加强元数据处理及与外部平台的无障碍链接和合作,对外部机构和团体发布数字资源。通过从单一内容来源中动态地发布所有格式(印刷品和电子)的商业信息,降低成本。同一个内容建设和标引完成以后,不必改变物理地址,由分布众多的接入点的链接,让期刊内容甚至是论文内的科学数据或者图表的价值实现最大化。

3)加强web2.0理念和技术作用,强调个性化服务和用户参与。通过提供准确、一致和相关的优质个性化文档,提高读者满意度,通过提供实时、定制的说明,提高出版、检索、下载和阅读效率。

4)重视统计分析等增值服务的开发。按需出版,根据客户的需求来确定出版内容和形式。通过简化创作、审查和发布过程,消除重复性工作。凡此种种,均有赖于生产环节的一次性信息生产。

5 结束语

流程优化是以全面质量管理和业务流程重构为基础的。由于数字化技术的发展和电子商务的出现,科技期刊出版面临着严峻挑战,要使内容生成有价值的产品并保有长效的生命周期实现可持续发展,必须从初始环节打造延续性的生产流程,以流程为导向创造出竞争优势。结构化排版是实现数字化出版的关键技术之一。基于XML的信息处理流程采用自动化信息处理技术高效地生产符合客户要求的内容产品,最大程度地削减信息应用成本和系统维护工作量,实现了结构化排版,对出版流程进行全方位改造,提供更为强大的信息服务(资源浏览、检索、个性化服务、增值服务等),为用户提供多样化的、有特色的选择。规范的数字出版流程的建立,XML文件以资源文件的新形式在出版体系和流程中理顺出版产业链的关系,不再仅仅是传统出版过程中的昙花[15]。

总之,结构化排版可以自主、灵活地,多渠道、跨媒体、多语种地为期刊内容资源的发布提供基础保障,加强数字资源挖掘整合,有效地再组合和利用,改变期刊以传统纸版发行、广告作为单一收入来源的经营结构,最终建立多元化、立体化的内容产品和增值业务经营体系,创造出新的数字经济增长点,推动我国科技期刊向着规模集群式和生产集约化方向发展[16]。

6 参考文献

[1] 杨根福. 从企业流程再造理论看印刷数字化工作流程[EB/OL]. [2010-01-02]. http://xxb.cnwest88.com/read.php?articid=1377&c=1

[2] 沈俊,缪淮扣. 用XML实现基于大型数据库的结构化排版[J]. 上海大学学报:自然科学版,2006,12(5):524-528

[3] 李亚青. 从学术期刊群数字化平台构建看高校学报数字化平台建设[J]. 中国科技期刊研究,2009,20(6):1084-1086

[4] 翟自洋,林昌东,林汉枫,等. 加入PubMed Central的实践

及其对期刊的积极影响[J]. 中国科技期刊研究,2007,18(5):761-765

[5] 周庆辉,凌昌全,白玉金,等. 美国《医学索引》选刊与收录方法及中国期刊收录现状[J]. 中西医结合学报,2005,?(?):70-78

[6] 查尔斯沃斯公司首页[EB/OL]. [2009-12-20]. http://www.charlesworth.com.cn/electronic~typesetting

[7] 陈志杰. LaTeX 入门与提高[M/OL]. 2 版. 北京:高等教育出版社,2006. http://wims.math.ecnu.edu.cn/ty/index-ty.html" \t "_self"

[8] 彭玲. 改进学术期刊出版流程 加快我国期刊数字化进程[J]. 数字图书馆论坛,2009,?(?):39-43

[9] 王强,王继成,武港山,等. Web 文档清洗系统中 HTML 解析器的开发[J]. 计算机应用研究,2002,?(?):54-57

[10] 邹振亚. 基于 XML 和工作流的期刊出版系统的设计与实现[EB/OL]. [2009-12-20]. http://c.wanfangdata.com.cn/periodical-jsjxtyy.aspx

[11] 《国家数字复合出版系统工程》标准研制工作方案:征求意见稿[EB/OL]. [2009-12-20]. http://www.hnppb.gov.cn/xwcb/pagesnew/sjxw.jsp?bh=999

[12] 刘金铭. HighWire 出版社评介:一个最好的学术期刊在线出版平台[J]. 中国科技期刊研究,2009,20(6):1012-1016

[13] 彭玲,张泽. 跨媒体出版的现状、问题及尝试[J]. 数字图书馆论坛,2008,?(2?):13-16

[14] 厉衍飞,刘培一. 我国学术期刊数字出版平台发展的分析与建议[J]. 中国科技期刊研究,2008,19(5):733-737

[15] 毛善锋,田杰,张莞. 预置 XML 标签 定制 DOI 元数据[J]. 科技与出版,2009(11):14-16

[16] Teri Tan. XML Publishing, the Way Forward. [EB/OL]. [2009-12-20]. http://www.publishersweekly.com/artictml?q=XML+Publishing%2C+the+Way+Forward

(2010-01-06 收稿;2010-02-26 修回)