CMA JATS 在中华医学会杂志社数字出版中的三年实践总结

■沈锡宾[1)] 李 鹏[1)] 刘 冰[2)] 姜永茂[2)]

收稿日期:2018-01-01
修回日期:2018-01-23

1)中华医学会杂志社新媒体部,北京东四西大街 42 号 100710
2)中华医学会杂志社,北京东四西大街 42 号 100710

摘 要 **【目的】**总结中华医学会杂志社的中华医学会期刊文档标签集(CMA JATS)3 年数字出版的实践经验,评估其在中文科技期刊数字出版和结构化排版中的作用,为制定中国科技期刊数字出版的数据标准提供参考。**【方法】**以中华医学会杂志社数据标准演进历程和 3 年实践经验为分析蓝本,阐释数据标准对于科技期刊数字出版的价值,总结经验,提出中文科技期刊数据标准的发展方向。**【结果】**CMA JATS 应用 3 年来,较好地满足了中华医学会杂志社对于数据整合的需求,累计加工期刊超过 150 种,9 万篇文献使用该标准进行数据加工并且实现多终端发布,基于该标准的数据接口可供第三方数据库应用获取论文数据。此外,CMA JATS 已经成为国家数字复合出版系统工程 XML 排版工具的内置标准之一,可以满足中文结构化排版的要求。**【结论】**CMA JATS 有效地整合了中华医学会杂志社的期刊资源,推动了中华医学会杂志社数字出版的发展。CMA JATS 也有能力成为中文科技期刊结构化排版的标准。

关键词 数据标准;科技期刊;XML;XML 排版
DOI:10.11946/cjstp.201801010003

近年来,中国科技期刊的数字出版取得了一定的发展[1-2],越来越多科技期刊的网站不再囿于 PDF 的全文展现方式,开始尝试全文使用超文本标记语言(HTML),HTML 极大地提升了阅读体验[3-5]。随之而来,人们意识到可扩展标记语言(XML)数据在数字出版中的价值[6]:一次加工,多元复用,并寻求将论文数据先加工为 XML 再网络发布[7-9]。这些实践印证了结构化数据对于科技期刊出版的重要性,但是笔者调研我国几家 XML 数据商的加工数据,发现 XML 数据质量参差不齐,有的加工商甚至对于数据标准没有清晰的理解(数据未发表),而已有研究对于科技期刊文献的标准探讨得还不是很充分[10],迄今也没有更多文献能够阐述中国科技期刊电子文档的规范性问题。

2014 年,中华医学会杂志社制定了国内第一个用于科技期刊全文数据标引的标准——中华医学会期刊论文标签集(CMA JATS)[11],至今已经历了 3 年的发展,标准自 0.1 版本升级到了 1.1 版本。本文总结 CMA JATS 3 年实践经验,深入地剖析该标准在中华医学会杂志社数字出版中的实践应用,阐明其在中华医学会杂志社数字出版转型中的价值,以此为中国科技期刊数字出版的发展提供参考。

1 CMA JATS 的制定和概述

1.1 CMA JATS 的渊源

进入 20 世纪,中华医学会杂志社着手制定数字出版转型策略,但是当时中国科技期刊界还没有一个被广泛认可且能应用于实践的行业标准来指导期刊数据的存储、交换和出版。2008 年,中华医学会杂志社组建了由 8 名学术编辑组成的数据标准研究小组,开展了对美国 NLM DTD 3.0 的研究工作,编写了 NLM DTD 3.0 的翻译文稿,共计 14.6 万余字,部分成果整理后发表在国内的科技期刊上[12-14]。这些研究成果使得国内学者洞悉了结构化数据对于科技出版的意义,了解到国外科技出版商数字出版的核心标准及其应用实践。基于对美国 NLM DTD 标准的研究,中华医学会杂志社着手制定适合中文科技期刊的文档标准。

科技论文作为人类知识传播的载体,必须置于开放的环境中,应当考虑国内国际间数据交换、存储和备份索引的需要,在出版标准的制定中遵循以下原则[10]。(1)包容性:适应多学科的内容描述,最大程

作者简介:沈锡宾(ORCID:0000-0002-7310-8157),硕士,副编审,E-mail:robin@cma.org.cn;李 鹏,学士,副编审,副主任;刘 冰,学士,编审,副社长;姜永茂,学士,编审,社长。

度地包容各出版商、图书馆、仓储数据库的需求。(2)开放性:考虑标准的演进和升级,为标准的发展留下足够的空间。(3)本土化:满足中文科技期刊的特性,合理地接受中文表达方式与国外期刊的不同之处。(4)国际化:在本土化的基础上考虑与国外知名数据库进行数据交换的需要。(5)可行性:寻求合适的内容精度,兼顾标准的实用性和经济性,不盲目追求内容标注的细颗粒度。

1.2 CMA JAST 制定流程

当时国内尚缺乏结构化数据标准方面的文献资料,所以只能参考国外科技期刊的成熟经验,摸索制定适合中文科技期刊标准的道路。CMA JAST 制定流程分为以下几个步骤:(1) 全面梳理 NLM DTD 3.0 标准,理解该标准对于文档内容描述以及还原的能力;(2)遴选有代表性的中文科技期刊,以此为依据手工制作 XML 数据;(3)在数据制作过程中,记录和整理 NLM DTD 对于中文期刊数据描述的不足,核查 NLM DTD 说明材料及其案例说明,确定上述问题是否超出 NLM DTD 可标记范围;(4)在 NLM DTD 基础上进行适当增删或者重构,形成第一个标准草案,制作 Schema 文档,并且以此制作第一个中文论文 XML 样例;(5)扩大期刊和论文类型范围测试标准的可行性,发现问题时再次修改标准,同时修订 XML 样例。

基于上述流程,经过半年多的实践,确定了 CMA JATS 的 0.1 版本[10]。相较于 NLM DTD,该标准有 2 个特点。(1)适度细颗粒度。CMA JATS 比 NLM DTD 删减了约 1/4 的实体。之所以删减部分实体,一方面由于某些实体对于数据标引的价值不大,性价比不高;另一方面,在多种标引方式中,只选择最合适的标引方式,以提升标引效率。(2)兼顾国际化与本土化。CMA JATS 继承了 NLM DTD 的大部分元素、属性和实体,但是为了适应中文科技期刊的特性,重构了一些元素,补充了小部分元素及其属性。

从 2006 年接触 NLM DTD[15],到 2008 年组建团队有计划地研制符合中国科技期刊特点的全文结构化标准文档,再到 2014 年 10 月 CMA JATS 0.1 版成形,发布于中华医学网,标准研发历时 6 年。CMA JATS 成为我国第一个由杂志社制定并且用于实践的全文数据标引标准。

2 CMA JATS 的应用场景

研制上线 3 年来,CMA JATS 成为整合中华医学会杂志社期刊数据的核心标准,贯穿于数字出版的全过程,在多个方面发挥了决定性的作用。

2.1 在资源结构化加工方面的作用

CMA JATS 1.0 标准为中华医学会系列论文的结构化数据加工提供了依据和准则,基于 CMA JATS 制定了期刊数字化技术要求和数据加工文档质量检测标准。期刊数字化技术要求从 7 个方面约定数据加工商提交加工数据的规格,分别为图像质量、PDF 文档质量、XML 文档质量、文档和数字对象唯一标识符命名规则、文档标准质量、文档内容质量和纸刊回库质量。中华医学会杂志社与数据加工商研制基于 CMA JATS 的数据加工流程。

经过全面评估,中华医学会杂志社从方正书版的文档入手,研制加工流水线。选择书版文档转换的一个原因是书版 FBD 文档保留了论文各构件的样式,通过这些样式信息可以还原对应的知识实体。最后通过手工标记的方式把作者、作者单位、参考文献等计算机无法准确识别的实体通过工具批处理。在上述流程研制成功后,扩大原始文档来源范围,尝试转换 PDF 或纸刊,或由 InDesign、飞腾、飞翔等排版软件输出的原始数据,确定方案是先将文档转化为可提取文字的 PDF,再通过 PDF 版面智能识别工具完成知识实体的识别,中华医学会杂志社最终构建了结构化数据加工完整流程。

在数据的交付过程中,数据加工商需要经过 2 步审查方可入库,第一步由计算机完成,XML 在交付前均需要通过 XMLSpy 进行 Schema 验证,无法验证的 XML 均视为不合格文档,此过程不仅保证了 XML 文档的良构性,还保证了 XML 中数据的合法性。第二步为人工检核,共三个阶段。第一阶段由数据管理员执行,数据在经过合法性验证后进入到资源池(入库)。入库时数据管理员首先对数据进行形式审查,然后将文档整期打包上传,此时资源管理系统再次检测数据的合法性。数据入库后系统解析 XML,生成可供检校的 HTML 页面,数据管理人员只需在此基础上进行检查。第二阶段由技术编辑和学术编辑检校,论文经过双重检查后,基本排除了所有问题,随即发布于正式环境,此后编辑部通过期刊发布平台查看论文,发现问题上报数据管理部门,数据管理部门审核后再次修订 XML 文档,再次上传和发布。第三阶段是年度审查,通过计算机分析,每种杂志选取 2 期杂志中图表最多的 2 篇论文作为审查样例,比照纸刊进行检查。发现问题后以报告文档形式向数据加工商通报,促进其提升数据加工质量。

通过双方紧密合作,现在已经形成了高效的工业化流水线,现刊可以在 3 个工作日返回加工数据,过刊数据可在 7 个工作日返回。截至 2017 年 12 月,以 CMA JATS 标准加工的医学期刊超过 150 种,其中中华医学会系列期刊 141 种,非中华医学会期刊 10 种,完成全文结构化数据超过 9.6 万篇,累计加工页码超过 30 万页,生产图表 22.6 万余张,构建了中华医学会杂志社的数字资源基础。

2.2 在资源管理平台和期刊发布平台中的作用

中华医学会杂志社期刊发布平台经历了 3 个版本:2015 年 10 月完成了第一版的研发,主要目标是实现结构化数据的在线发布,支持 CMA JATS 0.1 版本;2016 年完成了系统升级,支持 CMA JATS 1.0 版本,实现 XML 数据发布为手机版的 HTML;2017 年 10 月,在前期研发的基础上重构了系统架构,实现期刊发布系统与资源管理系统分离,增加了知识管理系统、用户/会员服务体系和知识付费体系。与 CMA JATS 标准有关的改进包括以下 2 个方面。(1)将资源管理系统(RMS)与期刊发布系统(CMS)分离,使其成为期刊发布系统的后台支持系统,提升了平台的整体稳定性。RMS 最重要的目标是建立一套以 CMA JATS 标准为基础的期刊资源数据管理系统。(2)引入知识管理系统(KMS)。作为知识服务的基础性工作之一,知识标引是提升论文价值的必要环节,该系统将 CMA JATS 标引数据通过计算机智能辅助技术实现知识标引。

2.3 在资源存储和交换中的作用

使用 XML 进行论文标引的初衷之一是利用 XML 跨设备和跨平台的特性,实现数据提供者间的交换。在 NLM DTD 问世前,科技出版商对于 XML 在数据存储和交换上的价值已经有了认识,制定了各自的 XML 标准用于资源的整合和交互,如今 XML 已经成为数据传输的主流方式[4]。中华医学会杂志社在制订数据标准时,也将 XML 在数据存储和交换方面的价值摆在极高的位置,要求 XML 数据不仅能够满足网络发布的需要,还能够用于数据交换。在实践中,本研究发现 CMA JATS 数据在数字资源的交换过程中起到关键作用。

2.3.1 通过 OAI 协议交换数据

CMA JATS 标准规范化的 XML 数据已经成为中华医学会杂志社对外传输期刊论文的重要文档格式,其中最主要的方式是将全文 XML 数据进行预处理,封装后通过 OAI-PMH 协议供第三方数据库或搜索引擎收割。

在数据资源入库的同时适当处理 XML 文件,保留前置部分<front>的大部分信息和少量后置部分<back>的信息。前置部分包括了期刊和文章的元数据信息以及版权信息等;后置部分保留了参考文献、文章的备注信息。经过上述处理后,将该文章关键元数据信息以 XML 形式存储于资源管理系统。

当资源发布系统响应第三方的数据传输请求时,系统将附上该 XML 文档的标准(删减版的 CMA JATS 标准),通过 OAI-PMH 协议对外发布。第三方数据库经过授权后从中华医学会杂志社收割论文数据。2016 年,该数据接口通过了中国科学院文献情报中心的技术联调,可供中国科学引文数据库(CSCD)获取论文题录信息以及参考文献数据。此外,中华医学会杂志社也通过该协议向百度学术提供数据。

2.3.2 向 PubMed 提交 XML 数据

PubMed 是国际知名的生物医学文摘数据库,自 2016 年年初,中华医学会杂志社将通过 CMA JATS 转化的数据向 PubMed 提交摘要 XML。因为 CMA JATS 的规格远超过 PubMed 的元数据要求,通过一套数据转化工具可以快速批量生产符合 PubMed 要求的数据。2016 年,PubMed 升级了数据标准(DTD PubMed 2.7)。与数字出版商协商后,中华医学会杂志社及时升级了数据转化工具,可以生产带中文摘要和英文关键词的数据,在与美国国立生物技术中心(NCBI)的工作人员沟通后,在 PubMed 数据库中首次展示了中文摘要信息。

上述过程得益于 CMA JATS 精准的结构化数据加工,在标准制订之初就考虑到为国内外科技数据库提供数据的可能性,不需要再返工便可以从既有的知识实体中抽取需要的内容,合成第三方数据库需要的数据文档,避免了为适应不同数据库的要求而重复加工数据的问题。

2.3.3 结构化数据的其他复用场景

有效发挥结构化数据价值的其他情况还包括在加工流程中导出文摘数据直接交给第三方数据库,比如中文生物医学期刊文献光盘数据库在数据加工后直接获得中华医学会杂志社的数据,避免了人力资源的重复投入。

此外,通过资源管理系统的数据接口,第三方应用或者网站可以实时从中获取数据,例如中华医学会系列期刊的微信平台可以通过接口获得题录和摘要信息,入库整理后在微官网平台展示。中华医学

会杂志社旗下的中华医学网和中华医学期刊网也可以通过接口获得期刊和论文的最新数据。

从上述实践可以看出,CMA JATS 数据除了可以在多个终端、多个平台发布外,更发挥了其数据复用的价值,实现了“一次制作,多元发布”,克服了数据重复制作导致的各种弊端,加快了数据的交换速度,提升了传播效率。未来,结构化文档复用的场景还可能有汇编出版、单行本出版、按需出版和重组发布等。

2.4 在结构化排版方面的作用

目前,中文科技期刊尚未进入结构化排版时代,基于排版文档的数据加工(俗称“后结构化”)仍是 XML 数据的主要来源方式,虽然可以满足数据存储、发布和交换的需求,但是后结构化加工模式相对落后,不符合出版发展趋势,造成资源的重复投入和损耗。为避免这些问题,采用结构化排版技术将是中国科技期刊数字出版发展的必由之路[16]。2015 年年初,作为《国家“十三五”时期文化发展改革规划纲要》和《新闻出版广播影视“十三五”科技发展规划》的重大项目之一,国家数字复合出版系统工程正式启动,其中第 11 包 XML 排版系统的目标是通过为 XML 结构化文档套用统一的版式模板,实现自动化排版、组刊及组稿,并在自动排版基础上实现版式的精细调整。该系统 2.0 版本于 2017 年 9 月 28 日通过了国家数字复合出版系统工程管理办公室的评审,取得了实质性成果[17]。中华医学会杂志社作为国家数字复合出版系统工程的试点单位,与项目承担单位的北京北大方正电子有限公司(以下简称“方正电子”)紧密协作,致力于现代出版技术的研发。

XML 排版系统研发伊始,方正电子即把 CMA JATS 1.0 作为中文科技期刊前结构化、自动化排版以及文档输出的基础标准之一。在第一阶段完成了版式模板制作工具、精调工具、XML 数据合成引擎、XML 输出引擎等的研发,之后基于该标准改进了 Word 智能结构化工具,可以完成 Word 文稿的智能结构化识别,转化成 CMA JATS 标准的 XML 文档。前结构化后的 XML 文档通过云端的排版引擎,套用排版模板,在合成引擎的渲染下完成自动排版过程,生成飞翔 FFX 文档和 PDF 文档。相较于传统排版技术,XML 排版技术大大缩短了初版文稿的排版时间,可以生成用于在线出版的多种电子文档。

实践证明,CMA JATS 定义的实体相对科学合理,可以满足中文科技期刊结构化排版的要求,有能力成为中文科技期刊结构化排版的文档标准。

3 CMA JATS 的未来

碎片化是未来知识服务的方向之一,碎片化的知识可以支撑个性化的知识产品,因此资源数据高度结构化是现代知识服务的必然需求[18]。未来 CMA JATS 极大提升知识服务转型的能力,主要体现在以下几个方面。

3.1 扩展内容标引广度和深度

CMA JATS 起初对于数学公式和化学公式的考虑有所不足,2017 年升级到 1.1 版本后支持数学公式的描述,以后还将保持开放性,采用其他结构化描记语言。例如随着期刊数据共享需求的进一步升温,数据共享相关的标准可能得到进一步强化。视频期刊的横空出世给科技工作者带来不同的体验[19],中华医学会杂志社也考虑创办纯视频类的医学期刊,而视频类论文将对现在以纸媒形式的数据标准造成很大的冲击。

因此,要重视数据标准的可扩展性,一方面寻求和引入成熟的第三方标记语言,增强对特殊组件的标引能力,另一方面扩展标记新兴实体的能力。

3.2 适应中文结构化排版需要

随着国家数字复合出版系统工程的顺利推进,2018 年 XML 排版系统将演进至 3.0 版本,届时结构化排版工具将应用于期刊出版实践,CMA JATS 将有真实环境来检验其实践价值。若在实践中与排版需求存在出入,可以在 CMA JATS 现行版本基础上开发子版本,制作专门为中文期刊结构化排版服务的版本。此外,还将密切关注 NISO JATS 的发展,取长补短,完善 CMA JATS 的标引能力。

4 结论

近年来,我国科技期刊界已经深入学习国际知名科技出版商对数据存储及结构化数据的应用,积极探索结构化数据对于中国科技期刊的实用价值。中华医学会杂志社也通过 10 余年的研究和 3 年多的实践,自行研制了 CMA JATS 标准,并且在多个领域取到了良好的效果。目前,中华医学会杂志社在结构化排版方面又与技术公司一道实现了该标准在期刊生产过程中的全流程应用。实践证明,该标准在知识实体标引的科学性、中文版式的适应性、标引颗粒度的适用性、标引实体的可扩展性等方面均表现优异,可以成为中文科技期刊结构化排版的文档标准。

纵观国际科技出版界 20 多年结构化文档标准的

发展历程,数据标准的建设不可能一蹴而就,只有广开言路,兼容并蓄、实地演练、不断精进才能求得好标准。CMA JATS 在未来的发展中,还将秉持开放的态度,欢迎业界同仁共享共用,共同在实践运用中推动 CMA JATS 发展升级,更好地为中国科技出版服务。

参考文献

[1] 程维红,任胜利,沈锡宾,等. 2011—2015 年中国科协科技期刊网站建设进展[J]. 中国科技期刊研究,2016,27(11):1156-1161.

[2] 程维红,任胜利,沈锡宾,等. 中国科协科技期刊数字出版及传播力建设[J]. 中国科技期刊研究,2014,25(3):340-345.

[3] 潘璇. 两种科技电子期刊平台的 XML 文档系统特点分析[J]. 中国科技期刊研究,2017,28(5):433-440.

[4] 刘雪梅,方曙,田丁. SGML/XML 在学术期刊电子出版中的应用与发展[J]. 中国科技期刊研究,2000,11(2):102-104.

[5] 白杰,杨爱臣. XML 结构化数字出版的特点与流程[J]. 出版广角,2015(5):28-31.

[6] 刘冰,游苏宁. 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[J]. 编辑学报,2010,22(3):262-266.

[7] 张光,白雨虹,刘文武. 利用 XML 技术提高期刊影响力的探索[J]. 中国科技期刊研究,2017,28(6):565-569.

[8] 王玥,南娟,刘谦,等. 基于 XML 的 InDesign 期刊排版文件标记与转换处理实践[J]. 中国科技期刊研究,2012,23(1):94-97.

[9] 侯修洲,黄延红. 基于 VBA 的 Word 文档 XML 结构化标记方法[J]. 编辑学报,2017,29(5):471-474.

[10] 沈锡宾,顾佳,韩静,等. 中国科技期刊文档格式标准化任重道远[J]. 编辑学报,2013,25(1):27-30.

[11] 沈锡宾,李鹏,王红剑,等. 中华医学会系列期刊全文电子文档交换和存储标准初探[J]. 中国科技期刊研究,2015,26(5):475-479.

[12] 沈锡宾,顾佳,包靖玲,等. 美国 NLM DTD 3.0 期刊存储和交换标签集中参考文献的标记解读[J]. 中国科技期刊研究,2013,24(2):233-237.

[13] 包靖玲,霍永丰,顾佳,等. 美国国立医学图书馆期刊文档标签集概述[J]. 中国科技期刊研究,2013,24(4):624-627.

[14] 包靖玲,李敬文,沈锡宾,等. 美国 NLM DTD 3.0 期刊存储和交换标签集中文章正文部分标记解读[J]. 中国科技期刊研究,2014,25(4):515-519.

[15] 沈锡宾,吕小东,郝秀原,等. PubMed Central 简介及其对期刊的评估和收录[J]. 中国科技期刊研究,2006,17(5):866-868.

[16] 扶文静,蒋湘莲,周泉. 基于 XML 的科技期刊排版生产流程再造及效益研究[J]. 价值工程,2016,35(13):89-91.

[17] 新华网. 推动数字化转型升级 国家数字复合出版系统工程取得重大突破[EB/OL]. (2017-05-22)[2018-01-23]. http://www.xinhuanet.com/politics/2017-05/22/c_129613225.htm.

[18] 刘红霞,沈锡宾. 重塑生产流程提升中国科技期刊知识服务的能力[J]. 科技与出版,2017(6):17-21.

作者贡献声明

沈锡宾:提出选题,制定标准,撰写论文;

李　鹏:制定标准,参与论文撰写;

刘　冰,姜永茂:设计论文框架,指导标准制定。

Three-years' case report of CMA JATS applying in digital publishing of Publishing House of Chinese Medical Association

SHEN Xibin[1)], **LI Peng**[1)], **LIU Bing**[2)], **JIANG Yongmao**[2)]

1) New Media Department, Publishing House of Chinese Medical Association, 42 Dongsi Xidajie, Beijing 100710, China

2) Publishing House of Chinese Medical Association, 42 Dongsi Xidajie, Beijing 100710, China

Abstract: [**Purposes**] This paper aims to summarize three-years' experience of Chinese Medical Association journal article tag set (CMA JATS) in digital publishing of Chinese Medical Association Publishing House (CMAPH), assess its value for XML typesetting of Chinese scientific journals, and provide standard references for digital publication of scientific journals in China. [**Methods**] Based on the history and three-years' practical experience, we analyzed the value of CMA JATS for scientific journals and indicated the development direction of CMA JATS. [**Findings**] The standard is well adapted to data integration of CMAPH. It has been used for digital processing for over 150 journals, and more than 90000 articles have been transformed to full-text XML documents and published online perfectly. The data interface based on CMA JATS can be used by third-party databases or applications to obtain article XML data. In addition, CMA JATS has become one of the built-in standards for Chinese XML typesetting tool, which validating its value in Chinese structured typesetting. [**Conclusions**] CMA JATS could effectively integrate article resources of CMAPH and promote CMAPH digital publishing. CMA JATS also could become the standard for structured XML typesetting of Chinese scientific journals.

Keywords: Data standard; Scientific journal; XML; XML typesetting

(本文责编:刘晶晶)