中国科技期刊文档格式标准化任重道远

沈锡宾[1)] 顾 佳[2)†] 包靖玲[3)] 韩 静[4)] 霍永丰[5)] 李 君[6)] 袁 庆[4)] 李敬文[7)]

1)《中华内科杂志》编辑部;2)首都医科大学附属北京朝阳医院科研处;3)《中华外科杂志》编辑部;4)《中华健康管理杂志》编辑部;5)《中华糖尿病杂志》编辑部;6)《英国医学杂志》(中文版)编辑部;7)《中华预防医学杂志》编辑部:100710,北京

摘 要 对于科技期刊的电子文档数据的存储和传输来说,文档格式的标准化是其基石。文章阐述 XML 作为科技期刊电子数据交换的统一格式的原因和历程,概述美国科技期刊电子文档化标准 NISO JATS 的发展历程及其组成和结构,介绍我国科技期刊文档标准化状况,并对制订有关标准提出建议。文章呼吁同行能在借鉴美国科技期刊全文标记实践的基础上,开展中文科技期刊文档格式标准化的研究,为中文科技期刊的按需出版、在线发布、全文数据库建设和数据共享打下基础。

关键词 科技期刊;电子文档;标准化;数字化;可扩展标记语言(XML)

A long way to standardization of electronic document of Chinese STM periodicals // SHEN Xibin, GU Jia, BAO Jingling, HAN Jing, HUO Yongfeng, LI Jun, YUAN Qing, LI Jinwen

Abstract For article archiving and interchange of STM periodicals, the functional work is to defin a suite of XML elements and attributes to describe the content and metadata of journal articles. The authors firstly explained why XML will be unified format for STM periodical electronic data interchange and its history. Then, the authors introduced the NISO Journal Article Tag Suite (NISO JATS) standard, including its history, development, contents and structure. At last, the authors described the situation of Chinese standardization for article archiving and interchange in STM periodicals and gave some suggestions for domestic standard establishment. Meanwhile, the authors called on colleagues carrying out such researches and discussions based on the American NISO JATS, for establishing the basis for publishing on demand, publishing online, full-text database construction and data sharing in Chinese STM periodicals.

Key words sci-tech periodical; electronic document; standardization; digitization; XML

First-author's address Editorial Board of Chinese Journal of Internal Medicine, 100710, Beijing, China

位于切萨皮克湾顶端西侧的巴尔的摩,是美国大西洋沿岸重要的海港城市。1904 年 2 月 7 日,一场世纪大火几乎将整座城市毁于一旦。当天,绸缎呢批发仓库里的一根烟头在 30 h 内烧毁了 90 条街区的1 526 栋建筑,累积损失超过 1 亿 5 000 万美元。令人痛心的是,火灾发生后,从哥伦比亚、费城和纽约等地闻讯赶来的消防队员只能在现场观火兴叹,因为他们的消防水龙头不能跟巴尔的摩的消火栓进行衔接。这件事件给美国政府一记沉痛的教训,也促成了美国国家标准与技术院(NIST)的诞生。在生活中,标准带给我们最大的裨益就是便利,其统一性让我们免去考虑很多匹配问题,比如手机的电源适配器、塑料的使用范围、灯泡的螺口等。

在期刊出版领域,尤其在文档数据的存储和传输方面,我们是否也真正拥有适于出版商的标准,以实现数据服务商和销售商的无缝收割? 国际上为此已作了很多努力,甚至形成了以某些标准为核心的软件生态系统(tool ecosystem)[1]。纵观国内,我们确实出台了很多标准来规范出版物的规格、开本、版式、装帧、编辑加工、校对等,但在科技期刊文档存储和传输方面却尚无相关标准,相关研究也鲜有报道。本文抛砖引玉,就我国科技期刊文档格式标准化的问题进行初步探讨,并呼吁业界同人能正视数据化过程中的标准化问题。

1 数据交换的统一格式:XML

涉及电子数据的存储和传输,我们必然要提到可扩展标记语言(XML)技术。狭义地说,XML 是一种使用标记(tag)的数据格式,既能让机器阅读,也能让人类阅读。利用这些标记就可以定义来自期刊论文、读者来信、发票、邮件、小说和其他任何形式的数据流的内容片断。

一般认为,期刊出版标记技术的应用起源于 20 世纪 90 年代晚期的 XML,但实际上可以追溯到 80 年代标准通用标记语言(SGML)的革命。期刊出版是首个成功应用标记技术的行业,国外期刊界为之积累了 30 多年的经验。在 20 世纪 80 年代早期,基于 SGML 的期刊生产、传播、显示和存储研发持续了一段时间;但其重点在于元数据的排版,针对期刊全文的标引工作并不是很活跃。最终由于 SGML 高昂的专业技术成本及较差的兼容性,在 90 年代万维网出现后被逐渐遗弃。

人们会问为何最终选择 XML 作为期刊数据交换的统一格式,而不是其他的文本格式或数据库? 解答这个问题就如同解答为何人们会从一般等价物中选择黄金作为货币一样,XML 的一些特性成就了它作为互联网及各行各业电子数据交换的统一格式的地位[1]。

† 通信作者

首先,作为轻量级的数据储存文件,XML 是一种以文本形式来描述的文件格式,可以方便地穿越防火墙,跨越不同平台进行正常的数据交换;其次,XML 使用有语义的标记,可以在互联网上进行数据交换保持原数据的意思和构造,又可保持在不同系统之间数据交换的灵活性;其三,用户可以自定义 XML 的标记名称和关系,也就是说可以根据 XML 的基本语法来进一步限定使用范围和文档格式,从而定义一种新的语言,这便是 XML 的可扩展性;其四,XML 使文档的内容和形式完全分离,这一特性为 XML 的应用带来了很大好处,出版商可以轻松地实现内容的多元化发布;最后,XML 支持多种编码,文档本身就包含了所使用编码的记录,方便多语言系统对其数据进行处理。

正如我们使用的自然语言,使用某种语言的人多了就会形成一套语言体系,外来人要融入该群体时,就会将此作为交流工具。同样,在特定的行业内可归纳出一套标记集合,即约定用一套特定的 XML 应用语言(通常使用 DTD(文档类型定义)、XSD(XML Schema 定义)或 RNG(RelaxNG)定义)作为交流工具是很有价值的。随着 XML 技术的不断成熟和实践经验的不断积累,涌现出大量与此相关的标准套件,让出版人最初的很多想法得以实现。比如在化学、数学公式、图片、动画和多媒体标记领域取得了很多代表性的成就,制定了包括 MathML(数学标记语言)、SVG(可缩放矢量图形)、CML(化学标记语言)、TecML(技术数据标记语言)、UnitsML(单位标记语言)和 STMML(科学、技术和医学标记语言)在内的一系列标准,使得 XML 标准家族不断壮大,形成规模,为全文标记奠定了夯实的基础。

2 美国国立医学图书馆期刊文档标记标准的发展历程

进入21 世纪后,XML 的崛起让国外期刊出版商逐渐抛弃以排版显示为目标的期刊生产模式,进入到以标记为中心的期刊生产模式。万维网出现后,XML 成为了各浏览器和平台实现数据流通与展示的利器。各软件供应商跨越不同的平台、操作系统和编程语言为 XML 创建了一个资源丰富的软件生态系统,从前期的稿件编辑、文档制作、排版、校对,到中期的出版、发布,再到后期的仓储、交换和二次文献整理,各个环节上均有相应的软件和服务作为支撑。设计这些工具的软件供应商已经认识到 XML 的使用最终会让出版重心转向语义丰富的"结构性"标记,而这种新标记又反过来促成出版商开发出多个交换模型,这些模型可以同时让数据在不同媒体中得到重复利用,比如印刷版、浏览器以及应用越来越普及的移动终端设备。

美国国立医学图书馆(NLM)下设的国立生物技术信息中心(NCBI)为能让出版商和数据库以一种通用的数据格式进行期刊内容的交流和存储,于 2003 年 4 月发布了第1 版 NLM DTD(定义了期刊数据 XML 标记的 DTD,又称 NLM JATS 标签集),之后迅速被学术出版界采纳。到目前为止,已有超过 3 000 种期刊,240 万余篇文献采用该 DTD 对其论文进行标注并被 PMC 数据库收录[3]。许多中小型出版社已经采用了 NLM DTD,很多大型的出版社也正准备借助 NLM DTD 来传输数据和内容。不仅如此,大部分大型期刊出版排版软件和服务提供商加快开发基于 NLM DTD 的应用和服务,并且愿意接受将 NLM DTD 作为文献标注的标准以实现内容的传输[4]。

在其他领域,比如对于数据聚合器,NLM DTD 已被广泛认可。目前,它已成为美国 Atypon Systems(http://www.atypon.com/)的主流 DTD,Ingenta(http://www.ingentaconnect.com/)和 HighWire Press 出版社(http://highwire.stanford.edu/)也将 NLM DTD 作为其全文内容交换的标准。当然,不得不提的是世界上最知名的 OA(开放存取)存储仓库,PubMed Central(PMC),一直是 NLM DTD 的制定者和执行者。在图书馆界,NLM DTD 也很受青睐,英国大不列颠图书馆(British Library)和美国国会图书馆(Library of Congress)已经宣布使用 NLM DTD 作为它们电子内容的储存标准。它也被 Portico 仓储库(http://www.portico.org/)采纳。

诚然,对 NLM JATS 有认知的读者会认为它是为 PMC 而生,但其实在创建之初,工作组就已广泛征求了整个科技界的意见,所以它是为所有的学术内容而设计的。在设计初期,它就兼顾了生命科学以外的科技期刊,比如经济学、物理、考古学、历史学等。在制定第 2 版时,NLM DTD 专家组甚至与包括 BioOne、Blackwell Science、Elsevier Science、HighWire Press、Nature 出版集团、University of Chicago 出版社、John Wiley & Sons 国际出版公司等出版商,Inera、Mulberry Technologies 软件开发商,美国物理学会、美国电子电机工程师学会等学会和 PMC 等 OA 数据库一起进行了商讨,所以 NLM JATS 可以适应整个科技界的研究成果的存储与展示。

3 各尽其用的4 套文档标签集

2008 年 11 月,NLM 更新了 NLM JATS 标签集到 3.0 版本,工作组对所有出版商有异议及与之前版本不兼容的问题进行修改,并决定不向后兼容。NLM

JATS 3.0 共包含 4 套标签集,每一套都有它的目的及使用范围。

1)期刊存储和交换标签集:为出版商与数据库、数据库与数据库、出版商与出版商之间传输和交换期刊数据提供了标准上的保障。

2)期刊出版标签集:比期刊存储和交换标签集对数据格式要求更加规范。它可以帮助出版商规范期刊数据以便在网页或纸质上展示。

3)论文创作标签集:为作者直接授权的内容提供了非常规范的格式。

4)NCBI 书籍标签集:用于标记在 NCBI Bookshelf 上的数据内容。

2011 年 3 月 30 日,NLM JATS 工作组将前 3 套标签集整理后作为 NISO JATS 0.4 提交给美国国家信息标准协会(NISO),经 NISO 期刊文章标准化标记工作组和 NISO 内容及征收管理专题委员会批准,发布了 NISO Z39.96x,JATS 0.4 作为试用的国家标准草案,并历经半年多的评审阶段,至 2011 年 9 月结束。2012 年 8 月 9 日,JATS 0.4 正式成为美国国家标准(ANSI/NISO Z39.96-2012)[5],至此,该标准成为全球电子期刊出版领域的第一个国家标准。

NISO JATS 标签集定义的文档是科技期刊上的顶级元素,比如论文(研究性和非研究性文章)、来信、述评、书评或产品评述等。每个文档由 1 个或多个部分组成,如果由多个部分组成,它们必须以以下顺序出现[6]:1)前置部分(front,必选);2)主体部分(body,可选);3)后置部分(back,可选);4)浮动部分(float,可选);5)评论(response)或次级论文(sub-article)(可选)。

4　中国期刊文档数据标准化尚处于初级阶段

尽管我国的数字出版产业发展势头强劲,但至今尚未形成健康、完整的产业链,而在盈利模式、版权保护、行业标准、品牌影响、渠道建设和人才队伍等诸多方面也还远不及传统出版业的成熟水平[7],尤其在科技期刊出版领域更显滞后[8-9]。目前,国内科技期刊数字出版的最大盈利者不是出版者自己,而是数字技术提供商和数据服务商,他们廉价购买内容提供商的资源,进行简单的数字处理后牟取最大的商业利益。这与国内出版社、杂志社等内容提供商的经营意识薄弱、经营能力不足有很大关系。有些出版者认识到数据出版的趋势及其将来带给自己的利益模式,但大部分出版者还在墨守成规中经营着不断萎缩的纸质产品,未能洞悉数字出版的未来和把握期刊发展的规律,未能在数字出版上取得实质性的进展,期刊电子文档存储和传输标准化的问题也就无从谈起。

笔者认为,国内的科技期刊出版人确实在很多方面面临着问题和挑战,比如技术、人才、工具、财力、流程的契合、编辑加工的规范化及编辑模式的改变,这些都是羁绊我们发展的现实障碍。在技术层面,国内科技出版界在标记出版这一领域几乎是空白,数字出版相关技术的引进和推广仍需要一段时间,而且缺少在此领域有丰富经验的复合型人才。此外,国内虽有某软件开发商在着力帮助中国出版界研发数字复合出版系统平台,以实现“知识标引、多重应用、一次制作、多元应用”的目的,但未有大规模应用的实践,多数期刊社仍处于观望或测试阶段。进一步说,要实现以标引为基础的数字化加工流程,需对原有流程进行颠覆性改变,国内很多小型的出版者还停留在内容的纸质展示方面,未必会痛下决心来重建流程。再加上出版模式改变中的编辑业务和相关知识的更新,对于某些编辑来说显得吃力,诸如上述挑战都在制约着我国科技期刊实现数字化转型的步伐。其中,贯彻始终的是期刊数据标引的标准问题,也是争议最多的一个问题。NISO JATS 及其前身近 10 年的发展历程告诉我们,数据标准不是一蹴而就的,只有不断进行探索和广泛听取各方意见才能完善。

5　关于期刊文档数据标准制订的建议

我国科技期刊电子文档格式标准的制订是一个需要有智慧和耐心来解决的问题,还有一段路程要走。在此之前,我们需要静心地去学习和借鉴国际科技期刊界的成功经验,探索出一套符合国际规范又立足本土的标准。标准的创建除考虑标准的一般共性,如可行性、经济性、实用性外,还有以下问题需要特别关注。

5.1　包容性　目前国内各数字加工商已有为网页展示而制定了数据标准;但作为竞争方,各家都是另起炉灶,很难实现数据的共通共享,而且基本上未在全文标记领域有所举措。出版软件开发商也在制订自己的 XML 标准;但其宗旨是为排版流程服务,所以不太适合当前的以标记为中心的出版生产模式。图书馆为方便馆藏文献管理,也在利用国际标准,但与期刊的数据储存和交换需求相去甚远;所以,应成立标准研讨小组,汇集各方意见,求得各家需求的最大公约数,制订出让各方接受的行业标准,最终将其上升为国家标准。

5.2　国际化　标准制订必须有国际化视野,在充分考虑我国期刊的情况下,多了解国际上成型的期刊元数据和大型数据库存储标准,让我们的元数据经标准化处理后能与这些数据库和存储机构进行无缝衔接。

美国的 DTD 定义经历了全球范围内大型出版商、数据提供商、仓储数据库、软件开发商、图书馆等近 10

年的实际考验,对于多语言的标引问题等也有一套成熟的解决方案,值得我们学习和借鉴。

5.3 开放性 标准之所以实用在于其不断在适应事物的发展规律,随着期刊出版活动的不断发展,需求也会产生变化;因此,保持标准的开放性就体现在不断对标准加以完善、补充和修订,实现标准的演进升级。

标准制订的过程一般是先在企业层次,进而上升为行业标准和国家标准。当行业层次的标准建立后,数字技术和服务提供商的重要性将会进一步削弱,内容提供者在整个数字化出版的产业链中的重要性才能真正体现,也非常有利于所有内容资源的整合和更好利用。若建立了国家标准,即便是推荐性的,对于规范和推动我国科技期刊的数字出版发展的作用也是无可估量的。国内的大型科技期刊出版商将会按照此标准制作、存储、发布和传输数据,而国内的服务提供商在获取版权的前提下无缝地收割数据并为读者服务。我们可以设想,通过基于此标准标引的数据聚集于某一聚合器的话,就可实现国家层面的大型文献数据库和应用服务。

6 结束语

XML作为元语言的功能和意义,科技期刊界有广泛的共识,但很多用户都是拿来主义者,习惯利用现成的"语言",很少会去考虑制作一套新的XML应用语言;所以,国内许多同人谈及电子文档标记规范,兴致不高,有人认为事不关己,有人认为遥遥无期,甚至有人挂个"内容为王"的口号作为挡箭牌,漠视国际科技期刊界发生的波澜壮阔的变革。这其实是反映了我们未能意识到"标准也是生产力"。所以,虽然我国数字出版产业发生了巨大变化,数字出版总产值也突破了1 000亿元,但传统出版的数字化转型依然没有取得实质性突破,与国外出版巨头相比还存在很大差距,短板之一便是标准[10]。标准之争其实质为市场之争,谁掌握了标准,谁就掌握了市场的主动权。

我们在此呼吁,国内同行能在借鉴美国NLM DTD的基础上,开展中文科技期刊文档格式标准化的研究与探讨,为中文科技期刊的按需出版、在线发布、全文数据库建设和数据共享奠定标准方面的基础,创造一个良好的行业氛围,酝酿并培育出中国的标准,道路可能曲折,但终究会成功。

7 参考文献

[1] Todd Carpenter. The value of standards in electronic content distribution:reflections on the adoption of NISO standards[J/OL]. Journal of Electronic Publishing,2011,14(1). DOI. [2012-09-19]. http://quod.lib.umich.edu/j/jep/3336451.0014.102?rgn=main;view=fulltext

[2] W3C Communications Team. XML in 10 points[EB/OL]. [2012-09-12]. http://www.w3.org/XML/1999/XML-in-10-points

[3] National Center for Biotechnology Information. PMC FAQs [EB/OL]. [2012-09-19]. http://www.ncbi.nlm.nih.gov/pmc/about/faq/

[4] INERA Inc. NLM DTD Resources[EB/OL]. [2012-09-19]. http://www.inera.com/nlmresources.shtml

[5] JATS:Journal Article Tag Suite[EB/OL]. [2012-09-12]. http://jats.niso.org

[6] National Center for Biotechnology Information. Journal Publishing Tag Set Tag Library version 3.0[EB/OL]. [2012-09-12]. http://dtd.nlm.nih.gov/publishing/tag-library

[7] 程维红,任胜利,路文如,等.我国科技期刊由传统出版向数字出版转型的对策建议[J].中国科技期刊研究,2011,22(4):467-474

[8] 王华菊,金丹,陈竹.科技期刊的数字化出版现状及问题探讨[J].编辑学报,2011,23(增刊1):9-11

[9] 丁岩,吴惠勤,龙秀芬,等.科技期刊数字化出版转型初探[J].编辑学报,2011,23(增刊1):3-6

[10] 刘成勇.推动数字出版进入高铁时代[J].出版参考,2011(21):1

(2012-09-18收稿;2012-09-30修回)

土地面积单位公顷的法定符号是 hm^2

国际计量委员会(CIPM)1996年建议,土地面积单位公亩(符号为a)、公顷(符号为ha)为暂时保留与SI并用的非SI单位。1 a=100 m^2,1 ha=100 a=10^4 m^2。在我国,公顷被选为国家法定单位,公亩则不是,理所当然不应采用词头"h"加非法定符号"a"构成公顷的符号"ha",而是另选定 hm^2 作为公顷的法定符号;因此,在中文科技书刊中,"公顷的符号就不能用ha,而应采用SI的十进倍数单位 hm^2"(见《量和单位国家标准实施指南》第24页)。由于1 hm^2 = 10^4 m^2 = 1 ha,因此,不必担心使用了 hm^2"别人会看不懂"。

尽管公亩不是我国法定单位,但如有需要,可用法定单位平方十米(符号为 dam^2)来替代,即1 dam^2 = 1 a。例如试验田大米的面积产量"1 000 kg/亩"可换算成"150 kg/a",进而用法定单位表示为"150 kg/dam^2"。

(郝 远)