# 基于 VBA 的 Word 文档 XML 结构化标记方法*

侯修洲 黄延红
《中国科学》杂志社,北京,100717

**摘 要** 利用 VBA 编程语言,针对科技期刊文章的特点,对 Word 文档实现了 XML 结构化标记,在全文标记的基础上,还对文档内在逻辑的连续性和一致性进行了自动校对。与传统出版模式相比,工作效率提高,编校错误减少。此方法已经应用于《中国科学》系列期刊,取得了良好效果。

**关键词** VBA 程序;XML 结构化标记;JATS;自动校对

**Method of XML marking in Word document by VBA program**//HOU Xiuzhou, HUANG Yanhong

**Abstract** In this paper we developed a method of XML marking in Word document by VBA program, and finished automatic proofreading for the document continuity and consistency. Compared to the traditional publishing mode, the work efficiency has been improved and the editing errors have been reduced. This method has been applied to the journals of Science China series, and performed well.

**Keywords** VBA program; XML marking; JATS; automatic proofreading

**Authors' address** Science China Press, 100717, Beijing, China

**DOI**:10.16811/j.cnki.1001-4314.2017.05.020

当前,不管是纸质版还是 PDF 电子版本的文章,已经不能适应碎片化阅读的需求,这就要求我们必须探索新型的出版形态。XML(extensible markup language)可扩展标记语言正是为了适应碎片化阅读的趋势,可以将一篇文章分解为若干个标记性段落,每一个标记性段落根据需要又可以分解为若干子项,通过标记与分解,信息的颗粒度会越来越细,从而将最有效的信息推荐给读者,以引起读者的注意,同时利用 XML 文件可以生成 ePub、HTML、PDF 等格式文件,真正做到一次生成多次发布,并且读者可以很方便地实现交互式索引、关联阅读和兴趣检索。

要将文档转换为 XML 文件,需要一个前提,即事先需要有一套文档各要素的识别方法,才能做到准确、快速的转换。笔者针对科技期刊的特点,尝试在 Word 文档环境中,并且该 Word 文档没有经过加工排版,利用 VBA(visual basic for applications)语言编写识别与标记程序,并嵌入 Word 工具选单中,实现一键即对 Word 元数据的标记工作。关于 VBA 语言环境、部署及具体案例应用可参考王玥等[1]的文章,也可以参考 VBA 的帮助文档。

*2015 年文化产业专项资金项目

当我们对文档有了识别方法后,还需要知道 XML 文件采用什么标准,具体涉及哪些标签,以便确定文档要素与 XML 标签的对应关系。经过调研,我们采用了美国国立医学图书馆(National Library of Medicine, NLM)颁布的 JATS 1.1(Journal Article Tag Suite, JATS)标准。

## 1 JATS 标签体系概述

JATS 1.1 标准版本是美国国立医学图书馆(NLM)于 2015 年 12 月发布的,最早的版本可以追溯到 2003 年,主要是作为文档存储、转换和 Web 数据交换的标准,供各出版商采用。其标签大体可以分为 <front> <body> <back> 3 大类:<front> 标签中主要包含文献的文摘信息标签,比如,刊名、ISSN、doi、学科、题名、作者、地址、E-mail、摘要、关键词、日期、年、卷、期、页码、基金及自定义标签等信息;<body> 标签中主要包含章节标题、段落、图、表、公式等信息标签;<back> 标签中主要包含致谢、参考文献和附录等信息标签。对这些标签的详细信息可以参考 JATS 的官方网页(http://jats.nlm.nih.gov/publishing/tag-library/1.1/index.html),JATS 官方网页不仅提供每个标签的详细解释、用途和示例介绍,而且提供了期刊的 XML 和 PDF 文档的 3 个样例,以方便使用者尽快熟悉 XML 标签体系。中文相关介绍可以参考包靖玲等[2]和沈锡宾等[3]的文章。

## 2 Word 文档结构化标记

在 Word 文档中标记,首先需要建立 Word 文档的样式集,然后将特定段落设定为特定样式,当所有段落都按规定标记相应样式后,即可完成 Word 样式与 XML 标签的映射。由于程序对文档的识别和标记不能保证绝对完整,特别是针对作者提交的未经编辑加工的原始稿件,还需要加工人员在标记程序运行后进行校对和修改;因此,在人工核对后再用专业软件转换为 XML 文件。

**2.1 Word 样式设计** 当已知 XML 标签名称,即可有针对性地为 Word 设计样式并命名。由于 XML 是一种不含具体文档格式的标记性语言,所以 Word 样式基本不设置具体格式,只用不同颜色区分不同样式。为了防

止设置样式时与文档中已经存在的样式冲突,特将新增样式名设置为英文名,并将原文中所有设置的外部英文样式删除后再添加新样式。英文样式名一方面与XML标签名接近,另一方面尽量用通俗易懂的英文单词,以方便后期加工人员处理文档时辨别样式和修改。

表1示出Word样式名与XML标签的对应关系(JATS 1.1标准,2015)。因为XML标签还附带中英文语言等相关属性,所以经常是1个标签可以对应映射多个Word样式。由于公式一般是OLE形式的外嵌式格式,比如MathType软件编辑的公式,不需要标记即可转换为MathML或LaTex公式。

一般新建样式我们用到的VBA命令为Set myStyle = ActiveDocument. Styles. Add(name:="abstract",Type:=1),该命令表示创建一个"abstract"段落样式,如Type:=2则表示设置的是字符样式。

**2.2 页眉、页脚和脚注的处理** 如果作者提交的稿件是按照期刊模版撰写的,或者我们在回溯过刊文档时,一般在页眉、页脚和脚注有一些信息是必须纳入到正文中来做标记的,本文则采用提取首页页眉、页脚相关信息,如年、月、卷、期、页码、栏目、专题等相关信息,然后按照相应样式写在正文之前。提取页眉的命令为Set myrange = ActiveDocument. Sections(1). Headers(wdHeaderFooterFirstPage). Range。在这条命令中,myrange即能得到首页页眉的段落;如果要得到首页页脚段落信息,则只需将上述命令行中的Headers替换为Footers即可。

**表1 Word样式名与XML标签的对应关系**

| 样式名 | 中文含义 | 对应XML标签 | 样式名 | 中文含义 | 对应XML标签 |
|---|---|---|---|---|---|
| publisher-name | 出版商 | < publisher-name > | abstract | 英文摘要 | < abstract > |
| short-publisher-name | 出版商缩写名 | < publisher-name > | abstract-zh | 中文摘要 | < trans-abstract > |
| journal-id | 刊物ID号 | < journal-id > | kwd-group | 英文关键词 | < kwd-group > |
| journal-title | 英文刊名 | < journal-title > | kwd-group-zh | 中文关键词 | < kwd-group > |
| journal-title-zh | 中文刊名 | < journal-title > | vol | 卷 | < volume > |
| abbrev-journal-title | 刊名缩写 | < abbrev-journal-title > | issue | 期 | < issue > |
| issn | ISSN号 | < issn > | first-page | 首页码 | < fpage > |
| cn | CN号 | < issn > | issue-title | 专题名 | < issue-title > |
| eissn | EISSN号 | < issn > | last-page | 尾页码 | < lpage > |
| article-id | 稿件号 | < article-id > | elocation-id | 文章编码 | < elocation-id > |
| article-id-doi | doi | < article-id > | copyright-holder | 版权拥有者 | < copyright-holder > |
| article-categories | 英文学科 | < subject > | copyright-year | 版权年 | < copyright-year > |
| article-categories-zh | 中文学科 | < subject > | funding-group | 基金信息 | < funding-group > |
| article-type | 英文栏目 | < article > | notes | 期刊主页 | < notes > |
| article-type-zh | 中文栏目 | < subject > | head-a | 一级标题 | < sec > |
| article-title | 英文题名 | < article-title > | head-b | 二级标题 | < sec > |
| article-title-zh | 中文题名 | < trans-title > | head-c | 三级标题 | < sec > |
| authors | 英文作者 | < contrib > | figure-caption | 英文图题 | < fig-group > |
| authors-zh | 中文作者 | < contrib > | figure-caption-zh | 中文图题 | < fig-group > |
| correspondence | 英文通信E-mail | < author-notes > | table-caption | 英文表题 | < table > |
| correspondence-zh | 中文通信E-mail | < author-notes > | table-caption-zh | 中文表题 | < table > |
| affiliation | 英文地址 | < aff > | footnote | 脚注 | < fn > |
| affiliation-zh | 中文地址 | < aff > | para | 正文段落 | < p > |
| pub-date | 出版日期 | < pub-date > | references | 参考文献 | < ref > |

对于脚注,可以用以下命令Set foot = ActiveDocument. Footnotes(1),表示foot是全文的第1个脚注,Set myrange = ActiveDocument. Range(foot. reference. Start, foot. reference. End)即能获取该脚注的段落信息。本文是将脚注的段落信息回写到正文引用的地方,并用字符样式进行标记,以方便后期转换。如果要获取每个脚注,可以用For Next命令来遍历文章中的所有脚注,然后依次处理。

**2.3 文摘信息识别与标记** 文摘信息主要对应XML标记中的Front标签所包含的子标签,一般是先按照期刊模版框架来识别。如果不是按照模版框架撰写的稿件,则默认第1个段落为题名,题名后面紧跟的段落为作者,其余信息则按照文字特征来判断。比如文字中类似*#####,[A-Z]*这样的则认为是英文单位地址,该字符串表示此段落中至少包含5位数字并且后面紧跟的是1个大写单词。具体命令为If i. Range Like *#####,[A-Z]* Then i. Style="affiliation",i表示定义的某个段落,i. Style命令表示这个段落标上"affiliation"样式。图1是< front >标签部分XML结构化标记结果示意图。

**2.4　章节、图题和表题识别与标记**　正文中如果是“1”“1.1”“1.1.1”开始的段落,我们分别定义为一级章节标题、二级章节标题和三级章节标题。比如要识别并标记一个二级标题,其命令为 If i. Range Like“[1-9].＊” Then i. Style =“head-b”,“head-b”即是一个二级标题样式。对于图题和表题,我们采用的规则是图片下方第 1 个非空白段落是图题,表格上方第 1 个非空白段落是表题,基本都能达到识别和标记的要求。

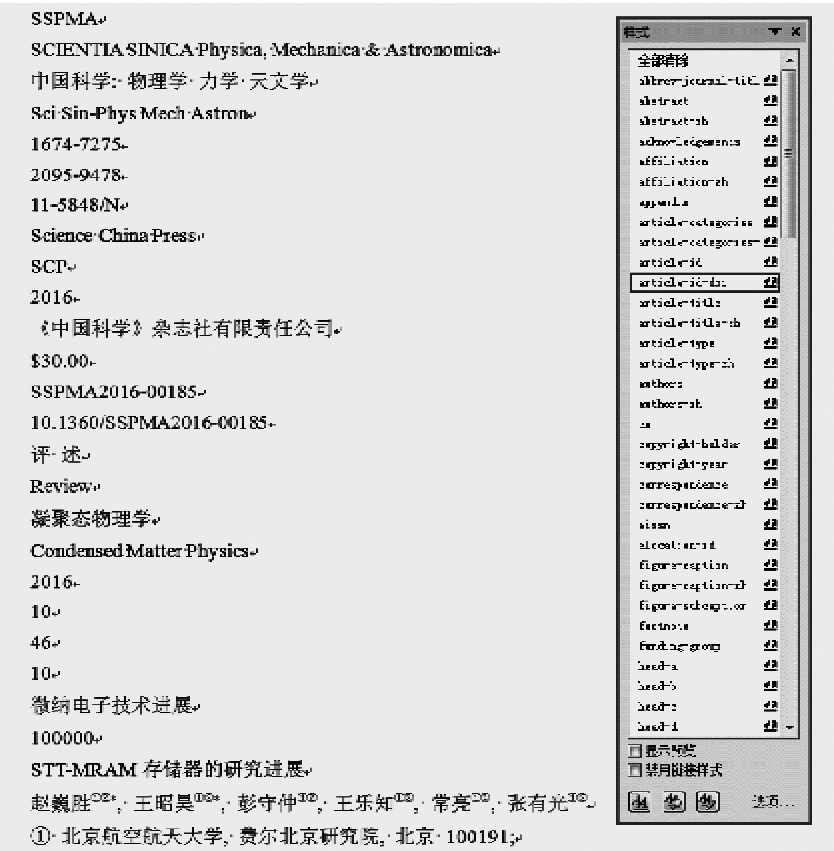


**图 1　Word 文档环境中 XML 结构化标记结果示意**

**2.5　参考文献识别、加工和结构化拆分**　对于参考文献,一般我们识别多个连续列表符号的为参考文献,如果程序没识别对,需要手动标记。

目前国内外对参考文献的处理一般是编辑先按照一定的格式进行加工,然后用程序来拆分文献的基本元数据,再用这些元数据去获取文献的 DOI 和 PubMed 等信息。这一处理流程从实践反馈来看有一定缺陷,主要是如果文献加工过程不规范,则会导致拆分错误,后期再从 XML 文件信息中去校对和修改,困难很大,严重阻碍生产的顺利进行。

笔者在前人的基础上,针对 Journal 类型的文献,提出了一种新思路,即先分析文献的体例特征,并将文献分为 5 大基本类型,按照分析结果去尝试解析文献的刊名、年、卷、首页码等信息,利用这 4 个数据,可以通过 CrossRef 的 API 接口获取到文献的 DOI 信息,再用这个 DOI 信息获取到 PubMed 和 ADS 数据库的 ID 信息,通过 DOI 和 ID 信息可以进一步去挖掘该文献的全部元数据,然后利用这些元数据来对我们的原始文献进行加工。具体方法和细节请参考侯修洲等[4-5]的文章。这个方法的优点是将标记和文献拆分、加工有机结合起来,加工人员只要一键即可完成所有文献信息的拆分。

## 3　全文逻辑连续性与一致性的自动校对

薛子俭等[6]曾经总结了一套科技论文分布编校法,该方法根据文档内容将编校过程分为“论文构架核查、归类加工、常规润色、整体性核对”4 步进行。虽然条理很清楚,也很严谨;但是全部都是需要编辑人工参与的,并且对编辑的能力和经验也是一大考验,一个轮次很难消除全部的编校错误。在 VBA 程序标记的过程中,我们发现全文在逻辑上存在内在的连续性和一致性,那么是否可以利用这些规律来实现文章的自动校对工作呢? 答案是肯定的。

比如文献顺序编码制,要求正文的文献引用必须按照顺序引用,不能漏引也不能跨文献序号引用,那么按照文献序号的连续性要求,可以对突然不连续的文献序号进行高亮标记,以提醒加工人员注意。依照这

一原则,图表序号、公式序号、章节序号的连续性校验同样有效地帮助了编辑加工人员。一致性校验则涉及著者-出版年制前后文的作者年信息匹配,如果人工校对,则是一个大量而又烦琐的工作,并且很难保证不出现一点疏漏或错误。程序化匹配则能非常好地保证高效匹配校验,并将不能匹配的"著者-出版年"引用信息用高亮标记提示出来。

对于中文期刊,程序还可以检查作者的中英文姓名和拼音是否一致,以及中英文地址邮编是否一致,如不一致则高亮标记。

对于文献,因为绝大多数Journal类型的文献都完成了拆分,如果2条参考文献的doi信息相同,则判断这2条文献是重复文献,程序会将这2条文献标上红色字体,如果拆分的文献中出现空标签,则该空标签会高亮标记。

通过全文逻辑连续性与一致性的自动校对,不仅减少了编校错误,而且对后期XML文档标签索引打下了很好的基础。这就减轻了加工人员的劳动量,提高了工作效率,也减少了审校的轮次,加快了出版速度。

## 4 结束语

本文按照JATS 1.1标准XML文件的要求,基本完成了对科技期刊论文的XML结构化标记工作,并在全文标记的基础上,对文章逻辑的连续性和一致性进行了校对,提高了工作效率,减少了编校错误。XML结构化标记是XML文件转换的第一步,也是非常关键的一步。之后我们引入了专业的XML转换工具软件,完成了标记后的Word文档向XML文件的转换,并通过XML文件和样式表发布了PDF和网站全文HTML版本,整个生产全部在我们最新的生产出版流程系统中完成,并已在《中国科学》系列期刊中实现了平稳运行。相比之前纯纸质版校样审校,现在的生产系统基本达到了无纸化、多地异地办公和跨平台无缝衔接,实现了从投审稿平台、XML排版生产和网上发布一站式流程。

## 5 参考文献

[1] 王玥,毛善锋,刘谦. Word文档中通过CrossRef自动查询与整合英文参考文献DOI的实践[J]. 中国科技期刊研究,2013,24(2):333

[2] 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

[3] 沈锡宾,顾佳,包婧玲,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

[4] 侯修洲,黄延红. 利用VBA程序和HTTPS协议获取参考文献的doi信息[J]. 编辑学报,2016,28(5):466

[5] 侯修洲,黄延红. 基于CrossRef数据库的参考文献自动加工及XML标引方法[J]. 编辑学报,2017,29(1):70

[6] 薛子俭,付利. 科技论文分步编校法及注意事项[J]. 中国科技期刊研究,2012,23(2):325

(2017-02-08收稿;2017-05-26修回)

# 第17届中国科技期刊青年编辑学术研讨会胜利召开

**本刊讯** 由中国科学技术期刊编辑学会主办,学会青年工作委员会和四川大学华西医院承办,华西医院期刊社协办的"第17届中国科技期刊青年编辑学术研讨会"于2017年7月27—29日在成都召开。300余位代表参会,围绕"影响力与学术生态"主题和"青年编辑实务"等热点问题进行了热烈讨论和广泛交流。会议收到论文108篇,评出优秀论文一等奖10篇,二等奖28篇,优秀奖26篇。

7月28日上午的开幕式由四川大学华西期刊社社长杜亮主持,青委会主任任延刚致开幕词。四川大学华西医院副院长程惊秋教授致欢迎词,学会副理事长栗延文编审、四川省新闻出版广电局彭佳副局长、国家新闻出版广电总局新闻报刊司原副司长张泽青莅会并讲话。各位领导对青委会的工作给予高度评价,殷切希望青年编辑牢记习近平同志的"广大科技工作者要把论文写在祖国的大地上,把科技成果应用在实现现代化的伟大事业中"的号召,努力传承精益求精的工匠精神,切实履行历史赋予青年人的使命,为将优秀科技论文发表在祖国期刊上而努力。出席开幕式的还有陈浩元、王亨君、姚希彤、肖宏、俞敏等。开幕式上,与会领导为获得优秀论文一等、二等奖的作者颁发了荣誉证书。

11位专家应邀在会上做了精彩的主题报告。中宣部出版局原副局长刘建生的《科学与出版》的报告,全面阐述了深入学习贯彻习近平总书记系列重要讲话精神和治国理政新思想新战略,把握正确政治方向和出版方向,以多出优秀作品为中心环节,着力加强内容建设、推进改革创新、完善出版管理,健全社会效益和经济效益相统一的体制机制,实现出版业持续繁荣发展,不断满足人民群众精神文化需求,推动我国加快从出版大国向出版强国迈进。邓洪新教授的《创办英文学术期刊的探索与实践》、张泽青的《关于学术期刊可持续发展的一些思考》、四川省新闻出版广电局报刊处处长邓志明的《四川省品牌期刊培育的探索与实践》、王亨君的《科技期刊在区域发展中的作用》、李幼平教授的《高质量证据生产的挑战驱动全程质控体系的创新与探索》等报告,以及在2个分论坛上15位青年编辑所做的紧密结合办刊实际的学术报告,都使与会者受益匪浅。

(蔡 巍)