

实战

预置 XML 标签 定制 DOI 元数据

东南大学学报(自然科学版)编辑部 毛善锋 中国科学技术信息研究所 田杰 张莼

摘要 为了实现期刊论文中文 DOI 元数据的批量自动提取,提出了向排版模板中预置元数据标签的解决方案。以方正书版文件为例,阐述了通过“不排”命令预置 XML 标签的具体方法,给出了采用自编工具软件提取 DOI 元数据和生成 XML 格式注册元数据文件的基本步骤,介绍了使用万方数据“中文 DOI 注册元数据转换与校验系统”完成元数据校验和网上提交的操作流程。

关键词 中文 DOI; 元数据; 自动提取; 网上提交

在 WWW 浏览器地址栏键入 <http://dx.doi.org/10.3969/j.issn.1001-0505.2009.03.001>, 敲一下回车键,直接呈现在电脑屏幕上的是《东南大学学报(自然科学版)》2009 年第 3 期第 1 篇文章的页面。其中 <http://dx.doi.org/> 是全球 DOI 解析服务器网址, 10.3969/j.issn.1001-0505.2009.03.001 是该文全球唯一的数字对象标志符——DOI 号。该页面包含了 DOI 号、文章标题、作者姓名和单位、摘要和关键词以及全文下载网址等 DOI 注册元数据信息。

要面向读者及时提供上述一站式解析和链接^[1]服务,就需要出版者(期刊编辑部)预先制作和提交包括 DOI 号在内的多项中文 DOI 注册元数据。为此,本文重点讨论 DOI 元数据自动提取技术,简要介绍 DOI 元数据的上传和注册流程。

1. DOI 元数据提取的技术瓶颈

由文献[2]可知,期刊论文中文 DOI 元数据包括期刊名称等 7 项期刊元数据和文章中、英文名称等 17 项论文元数据(含转页页码)。对大多数编辑部而言,目前这些 DOI 元数据的提取仍然以手工操作为主,即先从期刊论文电子版源文件中逐项摘取、复制元数据并粘贴到 Excel 模板“编辑部提供 DOI 注册元数据.xls”中,待整期期刊的

Excel 数据表制作完成之后,使用万方数据股份有限公司提供的“中文 DOI 注册元数据转换与校验系统”进行格式转换,最后再上传到 DOI 解析系统中。显然,这种手工摘取方式存在着质量和效率低、数据准确性和字符兼容性差以及进度明显滞后于期刊印刷版和数字版等问题。这些正是期刊论文中文 DOI 元数据提取的技术瓶颈。

针对上述问题,受《国家“十一五”时期文化发展规划纲要》重点项目“国家数字复合出版系统工程”和文献[3]中关于“一次制作、多元应用”设想的启示,以及查尔斯沃思(北京)信息服务有限公司 STM 杂志排版模式的启发,本文提出了将 XML 格式 DOI 注册元数据文件中的 XML 标签移植到排版模板中的解决方案。

2. 向排版模板中预置 XML 标签

XML(eXtensible Markup Language)即可扩展标记语言,它采用一系列简单的 XML 标签标记数据。XML 标签成对地出现,其基本格式形如“<author>李明</author>”,其中“<author>”是起始标签,“李明”是作者姓名,“</author>”是终止标签。

按照 DOI 解析系统的约定,将 XML 格式中文 DOI 注册元数据文件上传给中文 DOI 系统,是进行中文 DOI 注册和元数据存储的唯一通道,可见 XML 文件在 DOI 元数据与 DOI 解析系统之间扮演着“信使”的角色。关于其详细结构和内容,可参阅中文 DOI 网站上的 XML 文件实例。

批量自动提取 DOI 元数据的基础和捷径,是将 20 余项中文 DOI 元数据所对应的 XML 标签一次性地预置到排版模板中,为获得结构化的排版文件储备好 XML 标签,每个期刊编辑部或排版公司都能轻松地做到这一点。现以方正书版文件为例,详述 XML 标签的预置方法。

预置到排版模板中的外来 XML 标签,既不能影响原有的排版过程,又不得对排版结果特别是印刷版产生任何负面影响。方正书版系统的“不排”命令(BP)完全可以满足这个要求,其作用原理和命令格式也跟 XML 标签非常相像。

“不排”命令的格式为“[[BP(]]…[[BP]]”,意思是:内含左圆括号的“[[BP(]]”通知排版系统以下内容“不参与排版”;省略号“…”=“不参与排版”的具体内容,可以是文字、标签或其组合体;内含右圆括号的“[[BP]]”表明“不排”操作到此结束。灵活运用“不排”命令,巧妙组合“不排”内容,就能完整、准确地把定制 DOI 元数据所需要全部 XML 标签预置到排版模板中,并且既不增加排版工作量,又不影响排版结果。

在预置标签的过程中会遇到如下 2 种情况:

(1) 对印刷版中需要出现的元数据,必须采用 2 对“不排”命令即“[[BP(]]<起始标签>[[BP]]”元数据[[BP(]]<终止标签>[[BP]]”这样的标签格式,屏蔽掉 XML 起、止标签共 2 项内容,但将元数据保留下来。例如,本文第 1 段中 DOI 号的预置标签为[[BP(]]<doi>[[BP]]10.3969/j.issn.1001-

-0505.2009.03.001[[BP(]]</doi>[[BP]]]。

(2) 对印刷版中不需要出现的元数据,采用 1 对“不排”命令即“[[BP(]]<起始标签>元数据<终止标签>[[BP]]”,同时屏蔽掉 XML 起、止标签和元数据共 3 项内容。例如, CN 32-1178/N 所对应的预置标签为[[BP(]]<cn>32-1178</cn>[[BP]]。根据实践经验,为了使排版模板结构更加清晰,对内容相对固定的 7 项期刊元数据,以及其他不一定出现在印刷版中的元数据,可以全部采用这种标签格式,额外预置一组标签,并将它们集中摆放在排版模板的最前端。

XML 格式中文 DOI 注册元数据文件中,部分 XML 标签例如<cn>…</cn>可以被直接移植到排版模板中用作预置标签;部分双语元数据的标签同名,例如中、英文刊名的起、止标签都是<full_title>和</full_title>,其预置标签需要进行适当的变通处理。可以添加前缀“Chinese_”和“English_”,预置形如<Chinese_full_title>和<English_full_title>的过渡性 XML 标签,待生成 XML 格式 DOI 元数据文件时再还原为标准的 XML 标签。图 1 为包含预置标签的模板实例。

期刊中文名称: [[BP(]]<Chinese_journal_title>[[BP]]东南大学学报(自然科学版)[[BP(]]</Chinese_journal_title>[[BP]] 期刊英文名称: [[BP(]]</English_journal_title>[[BP]]JOURNAL OF SOUTHEAST UNIVERSITY (Natural Science Edition)[[BP(]]</English_journal_title>[[BP]] ISSN 和 CN: [[BP(]]<issn>1001-0505</issn>[[BP]][[BP(]]<cn>32-1178/N</cn>[[BP]] 年、卷、期: [[BP(]]<year>[[BP]]2009[[BP(]]</year>[[BP]] [[BP(]]<volume>[[BP]]39[[BP(]]</volume>[[BP]] [[BP(]]<issue>[[BP]]5[[BP(]]</issue>[[BP]] 本刊网站或第三方网站上的全文链接 URL: [[BP(]]<resource>http://service.wanfangdata.com.cn/File/download/Periodical-dndxb200903001.aspx</resource> [[BP]] (此处略去页眉排版命令) DOI 号: doi: [[BP(]]<doi>[[BP]]10.3969/j.issn.1001-0505.2009.03.001[[BP(]]</doi>[[BP]] 文章中文名称: [[BT1]] [[BP(]]<Chinese_article_title>[[BP]] …… [[BP(]]</Chinese_article_title>[[BP]] 作者中文名: [[BT2]] [[BP(]]<Chinese_author>[[BP]] …… [[BP(]]</Chinese_author>[[BP]] 作者单位中文名称: [[BT3]] [[BP(]]<Chinese_organization>[[BP]] …… [[HT]] [[BP(]]</Chinese_organization>[[BP]] 中文摘要: [[HTH]]摘要: [[HTF]] [[BP(]]<Chinese_abstract>[[BP]] …… [[BP(]]</Chinese_abstract>[[BP]] 中文关键词: [[HTH]]关键词: [[HTF]] [[BP(]]<Chinese_keywords>[[BP]] …… [[BP(]]</Chinese_keywords>[[BP]] 文章英文名称: [[WT3HZ]] [[JZ]] [[BP(]]<English_article_title>[[BP]] …… [[WT]] [[BP(]]</English_article_title>[[BP]] 作者英文名: [[HS2]] [[WT4]] B1 [[JZ]] [[BP(]]<English_author>[[BP]] …… [[BP(]]</English_author>[[BP]] 作者单位英文名称: [[WT6B1]] [[JZ]] [[BP(]]<English_organization>[[BP]] …… [[JZ]] [[BP(]]</English_organization>[[BP]] 英文摘要: [[WT5HZ]] Abstract: [[BP(]]<English_abstract>[[BP]] [[WTB1]] …… [[BP(]]</English_abstract>[[BP]] 英文关键词: [[WT5HZ]] Key words: [[BP(]]<English_keywords>[[BP]] [[WTB1]] [[ZK]] …… [[BP(]]</English_keywords>[[BP]] 起始页、结束页: [[HTH]]引文格式: [[HTSS]] …… [[J]]. 东南大学学报:自然科学版,2009,39(3): [[BP(]]<first_page>[[BP]]0000 [[BP(]]</first_page>[[BP]] - [[BP(]]<last_page>[[BP]]0000 [[BP(]]</last_page>[[BP]]

图 1 包含预置标签的排版模板实例(节选)

在包含预置标签的排版模板定型之后,每年只需在首期排版前更改一次卷号,每期只需在首

篇排版前更改一次期号等可变项目。采用这种模板进行排版,可以直接制作出包含 XML 标签和

DOI元数据素材的结构化、资源型排版文件,进而准确、方便地从中定制DOI元数据。

3. DOI元数据的自动提取和XML格式DOI注册元数据文件的自动生成

为检验预置标签方案的有效性和DOI元数据的定制效率,笔者采用Visual Basic 6.0语言编写了专用工具软件,其工作界面如图2所示。

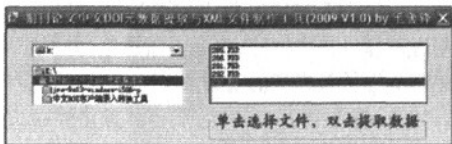


图2 自编工具软件的工作界面

在图2中,用鼠标双击包含预置标签的排版文件(例如293.FBD),软件随即开始逐篇查找文件中的预置标签,逐项解析出XML标签所标记的DOI元数据,清理无用信息,切分有用信息,初步完成DOI元数据的自动提取。接着,软件将元数据字符编码由方正书版的GBK内码转换成XML文件要求的UTF-8编码,并替换不兼容字符,再将DOI元数据和标准的XML标签等内容逐项写入同一个XML文件(例如293.XML)中,生成XML格式DOI注册元数据文件。每期期刊完成上述过程耗时5~10s,这在手工操作模式下是难以想象的。这种快至瞬间的处理速度,有助于实现期刊在线出版和元数据文件上传到DOI解析系统的同步化,且都先于印刷版,既方便读者随时访问、阅读和下载全文,又及时展示期刊的数字资源,赢得网上数字营销的先机。

4. DOI注册元数据文件的校验与提交

DOI元数据文件的校验包括脱机校验和联机校验2个层次。(1)脱机检验即XML文件语法正确性的一般性校验。用IE浏览器、Firefox浏览器或专业XML编辑软件XMLSpy打开上一节生成的XML文件,不出现语法错误提示信息即可。(2)联机校验包括XML文件结构完整性和字符编码正确性两大方面的严格校验。如图3所示,在万方数据股份有限公司提供的“中文DOI注册元数据转换与校验系统”中,“Schema格式检验”的结果显示,“293.XML”顺利出现在“校验正确

文件”的清单中。

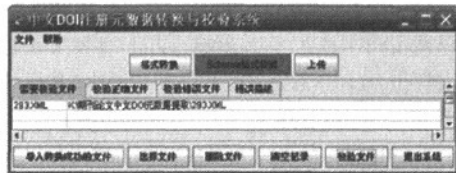


图3 中文DOI注册元数据转换与校验系统界面

接下来,将校验合格的XML文件通过图3系统上传到中文DOI元数据中心等待审核。审核无误即可将该XML文件正式注册到DOI解析系统中。至此完成期刊论文中文DOI元数据的自动提取、数据校验与网上注册的全部流程。

最后一个重要步骤,是及时备份排版文件、XML文件等原始数据和当前版本的元数据提取工具软件等。将来出现数字资源地址变化或其拥有者变更等情况时,网址元数据URL需要随之变更。这时就要利用原始数据对XML文件中的网址等信息进行更新,并再次将完整的XML文件上传到DOI解析系统中,覆盖原来的DOI元数据。也只有这样,才能把DOI的优势充分发挥出来,保证数字资源不会出现无效的网络链接。

结语

通过向排版模板中预置XML标签来实现DOI元数据的定制,是有选择地标记,有选择地提取,获得DOI元数据的质量和效率是普通数据挖掘或人工摘取方法无法比拟的。虽然本文做到的这一步与数字复合出版系统工程的远景目标“知识标引、多重应用、一次制作、多元应用”还相距甚远,但毕竟实实在在地跨出了完全不同于传统排版的一小步;排版过程可以由此演进为数字标引的过程,排版文件将以资源文件的新面貌出现,而不再仅仅是传统出版流程中的一朵昙花。

参考文献

- [1] 赵蕴华.国内数字期刊资源唯一标志符的应用研究.情报学报,2007,25(7):1018-1021.
- [2] 田杰.DOI在引文规范与链接中的作用.科技与出版,2008(12):61-62.
- [3] 田胜立.数字复合出版催生出版新业态.出版科学,2008,16(2):5-8,13.

预置XML标签定制DOI元数据

作者: [毛善锋](#), [田杰](#), [张莞](#)
作者单位: [毛善锋\(东南大学学报\(自然科学版\)编辑部\)](#), [田杰, 张莞\(中国科学技术信息研究所\)](#)
刊名: [科技与出版](#) 
英文刊名: [SCIENCE-TECHNOLOGY AND PUBLICATION](#)
年, 卷(期): 2009, (11)
引用次数: 0次

参考文献(3条)

1. [赵蕴华. 国内数字期刊资源唯一标志符的应用研究. 情报学报, 2007, 25\(7\): 1018-1021.](#)
2. [田杰. DOI在引文规范与链接中的作用. 科技与出版, 2008\(12\): 611-62.](#)
3. [田胜立. 数字复合出版催生出版新业态. 出版科学, 2008, 16\(2\): 5-8, 13.](#)

相似文献(0条)

本文链接: http://d.wanfangdata.com.cn/Periodical_kjycb200911004.aspx

下载时间: 2010年1月6日