# 网刊元数据自动提取和PDF文件自动分割实践
## ——以“中西医结合护理”网站为例

潘 新 滕 飞
(上海交通大学出版社,200030,上海)

**摘 要** 以《中西医结合护理》排版所用的方正书版文件为例,介绍用于圈定元数据字段的“准标签对”的选择技巧,以及fbd文件与html文件之间的字符兼容性和格式对等性的处理方法。以此为基础,可以方便地实现高质量网刊元数据的高效率自动提取,以及PDF文件的精准自动分割与转页合并。实践证明,对于特定期刊而言,上述工作是很容易自主完成的。

**关键词** 网刊;元数据;自动提取;PDF文件;自动分割-合并

**Practice on automatic extraction of metadata from Founder typesetting file and pages from PDF file for webzine: Website of *Nursing of Integrated Traditional Chinese and Western Medicine* as an example**//PAN Xin, TENG Fei

**Abstract** Taking the Founder typesetting file (fbd file) of *Nursing of Integrated Traditional Chinese and Western Medicine* as an example, this paper introduces the tips for selecting prospective tag-pairs to locate different metadata fields in fbd files, and the way to solve the problems of character compatibility and format equivalence between fbd file and HTML file. Thus, high-quality metadata can be automatically extracted from Founder typesetting files with high efficiency, and split-merge of pages from PDF file can be accurately realized. Practices have proven that all the above work can be easily finished for a particular journal.

**Keywords** webzine; metadata; automatic extraction; PDF file; automatic split-merge

**Authors' address** Shanghai Jiao Tong University Press, 200030, Shanghai, China

**DOI**:10.16811/j.cnki.1001-4314.2018.03.027

随着期刊采编系统的广泛采用,网刊也日益普及,提取和准备网刊必需的元数据和PDF全文已成为期刊社一项常规性的工作。关于从方正书版fbd文件中提取网刊元数据,近年陆续有文章发表出来[1-3]。文献[4]虽未直接涉及网刊元数据的提取,但其思路却是可以直接借鉴的。关于PDF文件的自动分割与转页自动合并,尚未见有文章探讨。

本文基于《中西医结合护理》的实践,分享进行方正书版文件的网刊元数据自动提取以及PDF文件自动分割、合并等工作的一些体会。

## 1 选择用于圈定元数据字段的“准标签对”

方正书版fbd文件是纯文本格式的文件,只有少量的文字和排版命令符号在其他系统或平台上不能正常显示。从方正书版fbd文件中提取网刊元数据,需要解决好以下4个方面问题:1)替换不兼容的字符,保证内容对等;2)将方正书版中的黑体、斜体、黑斜体、上下标和简单的数学公式等所涉及的排版命令,置换成对应的html网页的标签对,保证格式对等;3)准确定位、截取和重组fbd文件中的各个元数据字段,保证信息完整;4)根据所用网刊系统(玛格泰克、勤云、三才等)的个性化要求,对各元数据字段进行必要的编排,保证准确显示。这4个方面的数据规范,同样适用于处理好其他来源的网刊元数据。

文献[4]提出的在方正fbd文件排版模板中一次性预置一套确定性标签的方案,经过其编辑部近10年的实践检验,已能够保证一键完成高质量DOI注册元数据和网刊元数据的自动提取。实际上,这种方案适用于所有采用方正书版排版的期刊。

对于排版模板中不含文献[4]中那样的预置标签的方正fbd自有文档,可以借鉴该文献的思路,把fbd文档中已经存在的一些相对固定的排版命令甚至文字等作为“预置标签”来使用,本文称之为“准标签”。例如,可以把《中西医结合护理》一级标题的字体、字号命令“〖HT2H〗”认定为该级标题的起始标签,把其结束字体命令“〖HT〗”认定为标题的结束标签。对提取网刊元数据来说,上述“〖HT2H〗……〖HT〗”标签对(以下简称为“准标签对”)的作用,完全等同于文献[4]中的“〖BP(〗<Chinese_article_title>〖BP)〗……〖BP(〗</Chinese_article_title>〖BP)〗”标签对;二者都能准确无误地圈定一级标题,并支持准确提取该级标题的全部内容。依此类推,可从现有fbd文件的排版命令或特定文字、标点组合(例如“摘要:”)中,为每个元数据字段找到相应的标签对,并且这种标签对通常都是唯一的。即使个别甚至多个元数据字段前后暂时缺少这样的“准标签对”,也可同排版员一道,通过简单的构思,为每个元数据字段“制造”出具有唯一性的“准标签对”,用于准确圈定这些元数据字段。以后排版时统一采用这组约定好的“准标签对”即可。

总之,对于《中西医结合护理》这本特定期刊来说,由于其开本和版式设置、字体和字号配置等都是相对稳定的,因此就地取材,找到提取网刊元数据所需的

部分标签对,再对必需的排版命令进行适当的重组,从而形成了一套独特的“准标签对”体系。最后将这些“准标签对”写入文献[4]作者毛善锋提供的VBA共享程序——Word版的“Sub《中西医结合护理》玛格泰克网刊元数据自动提取. docm”中,形成类似于表1那样的“准标签对”清单。对docm文件中现成的VBA代码进行必要的微调后,即可进行网刊元数据的自动提取,生成符合网刊系统规范要求的Excel数据表格。

**表1　《中西医结合护理》元数据字段及fbd文件中的“准标签对”(节选)**

| 序号 | 元数据字段名称 | 起始标签 | 结束标签 |
|---|---|---|---|
| 0 | 文章边界 | 〖HT5〗(用于在内存中将整期fbd文件切成单篇) | |
| 1 | 年 | 2096-0867( | ) |
| 2 | 卷 | 年第 | 卷第 |
| 3 | 期 | 卷第 | 期 |
| 4 | DOI号 | DOI: | 〖 |
| 5 | 文章栏目 | 〖JY(〗 | 〖JY)〗 |
| 6 | 中文大标题 | 〖HT2”H〗 | 〖HT |
| 7 | 中文作者姓名 | 〖HT4”K〗 | 〖HT |
| 8 | 中文工作单位 | 〖HT5”SS〗( | 〖HT |
| 9 | 中文摘要 | 摘要: | 关键词: |
| 10 | 中文关键词 | 关键词: | 中图分类号 |
| … | … | … | … |
| 25 | 参考文献 | 〗参考文献 | 〖JY〗 |

## 2　网刊元数据的自动提取

VBA是Visual Basic的一种宏语言,是微软开发出来在其桌面应用程序中执行通用的自动化任务的编程语言,寄生于Word等Office系列软件中。对设计者来说,Word等就是天然的VBA开发环境;对使用者来说,Word等更是天然的VBA使用平台,只要电脑中安装了完整版的Word等软件就行。本文仅介绍通过VBA实现网刊元数据自动提取的2个重点环节。

**2.1　fbd文件的字符兼容性和格式对等性处理**　在html格式的网刊中,字符的粗体、斜体和上下标等格式信息,需要采用html自己的“排版命令”进行定义,例如,粗体的“目的”“方法”表示为“<B>目的</B>”“<B>方法</B>”;“P<0.01”则表示为“<I>P</I><0.01”,其中的“<”代表半角的小于号“<”。通过将方正排版命令转换为html“排版命令”,即可得到形如“<B>目的</B>……<B>方法</B>……<B>结果</B>……(<I>P</I><0.01),……(<I>r</I>=-0.209,<I>P</I><0.05)。<B>结论</B>…”这样的html版摘要文本,在html页面中可以看到跟方正大样对等的显示效果。为此,通过“txt=文件标准化(txt)”语句调用文献[4]中设计的fbd文件标准化处理函数,完成不兼容字符的替换;通过“txt=斜体粗体上下标和简单数学公式(txt)”语句调用文献[4]中所采用的格式兼容性处理函数,完成fbd文件与html文件中相关格式的对等性转换。

**2.2　网刊元数据的提取**　首先,通过语句“txt = 回收转页内容(txt, "文章编号:")”调用下转页内容回收函数。然后,在内存中用“fbd_txt = split(txt,"〖HT5〗")”语句将整期fbd文件切分为单篇。

接下来,读取“Sub《中西医结合护理(中英文)》玛格泰克网刊元数据自动提取. docm”表格中的“准标签对”,每次将一个“准标签对”的起始标签赋值给变量tag_start,将结束标签赋值给变量tag_end,通过“txt1=Metadata(fbd_txt(i), tag_start, tag_end)”这样的语句调用公共提取函数,即可完成一个元数据字段的提取。依此类推,即可一次性完成整篇、整期的元数据提取,并一次性将所有数据逐项写入Excel表中,生成网刊发布所必需的元数据。该过程只需让VBA程序自动完成各元数据字段的定位、截取和内容整理,并不需要复杂的理论和算法,且其总运行时间以秒为单位。

## 3　PDF文件的自动分割与转页合并

除了使用网上可找到的一些独立运行的PDF文件分割、合并工具(例如PDF Split Merge)之外,还可设计自己期刊专用的PDF文件自主分割、合并工具。

在上一章提取网刊元数据阶段,已经提取到每篇文章的起、止页码和转页页码(如果有)。用VBA读取这些页码数据,并从后台调用Adobe Acrobat专业版的提取页面功能,即可瞬间完成整期PDF文件的分割与转页合并。程序运行的主要步骤为:1)自动创建一个空白PDF文件,通过VBA代码“PDDocTarget.InsertPages(-1, PDF总文件, 起始页序号-1, 本篇页数,