学术期刊元数据处理例证分析*

郭晓亮[1,2)] 景 勇[1)] 张 璐[1)] 吉海涛[1)] 郭雨梅[1)] 黄仲一[3)†]

1)《沈阳工业大学学报》编辑部,110023,沈阳;2)西北大学西北联大与大学文化研究院,710127,西安;

3)《华侨大学学报(自然科学版)》编辑部,362011,福建泉州

摘 要 随着学术期刊数字出版的发展,元数据处理与应用日益受到重视。梳理近10年期刊元数据研究进展,认为其可分为综述研究、建模研究、采集与应用研究、转换与标准化研究4类,目前与学术期刊紧密结合的元数据研究,个性化、定制化提取方式探索,低门槛且结合日常工作的应用仍有待加强。筛选编辑常见元数据疑难问题,结合实例给出简捷高效的处理方法,旨在解决元数据应用的实际问题。

关键词 学术期刊;数字出版;元数据;数据整理;数据转换;例证分析

Example analysis on metadata processing in academic journals

// GUO Xiaoliang, JING Yong, ZHANG Lu, JI Haitao, GUO Yumei, HUANG Zhongyi

Abstract Along with the development of digital publication of academic journals, more and more attention is paid to the processing and implementation of metadata. Based on summarizing the research development of metadata of academic journals in recent 10 years, we divided metadata into 4 categories: summary research, modeling research, research on collection and application, and research on conversion and standardization. At present, it still need to enhance the research on metadata closely combined with academic journals, the exploration on personalized and customized extraction methods, and the application of low threshold and combined with daily work. Also, we selected common problems of metadata faced by editors, and put forward simple and efficient processing methods combed with examples, in order to solve practical problems in metadata application.

Keywords academic journals; digital publication; metadata; data collation; data conversion; example analysis

First-author's address Editorial Office of Journal of Shenyang University of Technology, 110023, Shenyang, China

DOI:10.16811/j.cnki.1001-4314.2019.06.017

元数据是描述数据仓库及其环境的数据,如存储位置、存储类型、转换规则、基本属性等,在期刊领域一般引申为对学术论文及其关联信息的结构化描述,属于基本数据单位。学术期刊网络化、数字化、智能检索等离不开元数据,其成为近年研究热点之一。以“元数据”“期刊”为关键词在中国知网检索,近10年文献342篇,其中46篇高相关论文可分为以下4类。

1)元数据综述。1996年起研究逐渐增加,2007年达到高峰并逐渐走向成熟[1],分理论研究、应用研究、互操作研究3个维度[2],并梳理了元数据研究进展[3-5]。但我国研究能见度和国际影响力较低,元数据主要在图书馆、档案馆,以及科学信息描述与数据管理和搜索引擎中应用[6]。

2)元数据建模。面向语义出版的数字资源聚合模型[7];基于生命周期、质量维度和影响要素对其质量管理建模[8];结合NSTL规范探讨大数据环境下其统一性、模块化、细粒度、标识性、关联性特点[9]等。

3)元数据采集与应用。利用网络爬虫获取元数据[10-12]、利用排版文件或PDF提取元数据[13-14]、通过JAVA或VBA等编程实现批量采集[15-16]等。应用方面,主要关注数据仓库建设[17]及元数据在数字图书馆、期刊数字内容挖掘等场景下的应用[18-19]。

4)元数据转换与标准化。通过科学数据与其互操作实现统一检索[20],对其进行转换与集成[21],通过数据清洗等技术措施[22-23]推进其标准化[24]等。

从事图书情报研究的学者对元数据关注更多,早期技术类和计算机类期刊发文较多,元数据采集研究近年来相对较为集中。相对于艰深的模型、算法,人们更关注的是如何更有效地利用元数据及其衍生品更好地把握前沿、提升效率、加强传播;因而,除了宏观媒体融合研究[25]外,笔者也试图解决实际技术问题[26],但与学术期刊紧密结合的元数据研究,个性化、定制化、智能化提取方案,低门槛且高频度的应用仍有待加强。本文以玛格泰克网络版采编系统和Excel 2007为例[27-28],提出具体方案。

1 规避科学计数法与“'”号处理

采编系统导出的xls元数据中的身份证号、银行卡号等经常会“变身”,即超过15位的数字后几位会变成“0”且无法恢复。这是因为软件默认将此类数字以科学计数法表示。

解决这一问题的方法主要有2个。一是先选中要粘贴的目标单元格后单击鼠标右键,选择快捷选单里

*辽宁省社会科学规划基金重点项目(L17AXW002);中国高校科技期刊研究会专项基金课题资助项目(CUJS-QN-2018-034);全国理工农医院校社科学报联络中心基金资助项目(LGNY18A2)。

†通信作者

“设置单元格格式”,并在“数字”标签里选择“文本”一项,这样将数字粘贴过来之后即会识别为文本,不会变化。但从采编系统里导出的元数据会自动在数字前加“’”号以保持文本格式,导致生成报表时存在无效字符,这就要采用第2种方法——通过数据分列清除“’”号并保持15位以上数字以文本格式正确存储。如图1所示,采编系统导出的原始数据前面带有“’”号或“◤”以标记其文本格式。图1及后文中各身份证号等个人信息均为虚拟。

金额	身份证
480.0	210628199007184879
320.0	210103196404111817
320.0	219054198901152631
320.0	219244197609132711
240.0	210103196404111817

图1 采编系统导出xls元数据示例

选中要处理的列,选择“数据”标签下的“分列”工具,如图2所示。分列处理前要先删去表头只保留数据列,因为软件不能对合并的单元格进行分列。

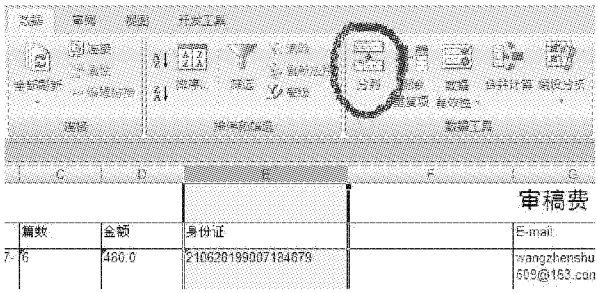

图2 “数据”标签“分列”工具

打开“文本分列向导”后,保持默认“分隔符号”,单击“下一步”按图3设置参数,将文本识别符号选择为“’”并查看预览区显示是否正确,无误后点击“下一步”按图4设置数据输出格式为“文本”。

图3 分列参数设置

图4 数据输出格式设置

最后点击“完成”即可去掉元数据的文本标志符“’”,使证件号等正确显示。由于能自动识别,此时相应单元格左上角仍有“◤”表示数字被储存为文本,但将这些元数据复制到其他软件后不会显示任何前缀。

2 带“◤”的元数据转换为数值型

用采编系统导出的xls元数据计算稿酬时,由于其以文本形式存储并以“◤”标记(如图1所示),无法进行求和等运算,此时就要将其转换为数值型。如图5所示,选中待转换单元格,点击左上角提示标记并在下拉菜单中选择“转换为数字”,即可完成转换。

图5 将文本转换为数字

3 同结构元数据表改扩展名与记录合并

在处理作者信息时,有时需要对导出的xls元数据进行记录合并,如邮寄样刊时需先查询单条作者信息,之后把每条作者信息作为一条记录存储到同一个表中打印邮签。手工复制粘贴在记录较多的情况下耗时耗

力,运用批处理命令和宏编程则可使问题迎刃而解。

在采编系统里查询需要的作者信息,导出时以“1,2,…,n”命名并放在同一个文件夹中备用。由于浏览器兼容性问题,导出的 xls 文件扩展名可能变成“.action”,此时用 Windows 自带“记事本”程序输入通配符重命名语句:

```
ren *.action *.xls
```

使用组合键“Ctrl+S”保存,并按图 6 所示保存为“.bat”批处理文件,将该文件放在存有单条作者信息的文件夹中,双击运行后原有“.action”文件即可全部改为“.xls”格式。

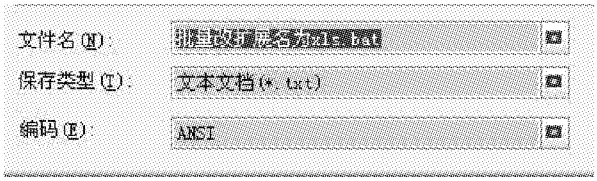

图 6　改扩展名为 xls 批处理文件命名

接下来通过一段宏代码实现同结构元数据表下的单条记录合并:

```
Sub 合并本文件夹下全部 Excel 表格()
  Dim MyPath, MyName, AWbName
  Dim Wb As Workbook, WbN As String
  Dim G As Long
  Dim Num As Long
  Dim BOX As String
Application.ScreenUpdating = False
MyPath = ActiveWorkbook.Path
MyName = Dir(MyPath & "" & "*.xls")
AWbName = ActiveWorkbook.Name
Num = 0
  Do While MyName <> ""
  If MyName <> AWbName Then
  Set Wb = Workbooks.Open(MyPath & "" & MyName)
Num = Num + 1
  With Workbooks(1).ActiveSheet
.Cells(.Range("B65536").End(xlUp).Row + 2, 1) = Left(MyName, Len(MyName) - 4)
  For G = 1 To Sheets.Count
Wb.Sheets(G).UsedRange.Copy.Cells(.Range("B65536").End(xlUp).Row + 1, 2)  '控制合并时每个表格取的行数,应以行数最多的表格为准,避免记录丢失。
Next
WbN = WbN & Chr(13) & Wb.Name
Wb.Close False
End With
End If
MyName = Dir
  Loop
Range("B1").Select
Application.ScreenUpdating = True
MsgBox "共合并了" & Num & "个文件中的全部工作表,如下所示:" & Chr(13) & WbN, vbInformation, "合并完成"
End Sub
```

新建 Excel 文件,选择“开发工具”标签,点击“控件”中的“查看代码”,将上述代码粘贴进输入框后保存,命名为“表格合并—打开时启用宏—开发工具—宏—执行.xls”,以提示不熟悉的使用者。打开文件时启用宏如图 7 所示。

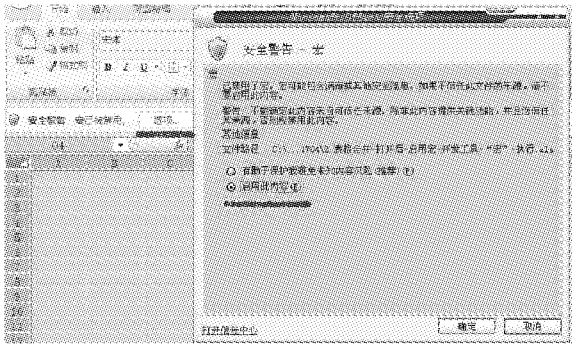

图 7　启用 Excel 文件中的宏

将该 Excel 文件放在存储单条作者信息的文件夹中,打开后按文件名提示的步骤操作即可得到合并好的表格。(上述批处理文件和宏文件可至 https://pan.baidu.com/s/1CB1Y9fF1syKB9tUcZ8PqAA 下载后根据需要修改使用。)

4　元数据格式整理与打印设置

在其他应用系统中进行批量查询时常需将导出的多个身份证号、电话号等以“,”“;”等符号分隔。此时,先将单条作者信息按上文方法合并为同一个表中的多条记录,再用上文的分列法去除“’”号;选中该列数据粘贴进“记事本”文件,以去除框线只保留元数据;用组合键“Ctrl+A”全选并粘贴进一个新建 Word 文档,形成每列 1 个号码的格式。在 Word 中点击“开始”标签下的“替换”,在打开的“查找和替换”浮窗中点击左下角“更多(M)>>”按钮,并在“特殊格式(E)”中选择“段落标记(P)”,之后在“替换为(T)”文本框中输

入“,”或指定分隔符,最后点击“全部替换(A)”即可将元数据处理为所需的查询字符串,如图8所示。

图8 替换参数设定

录入网刊元数据时,需要对“参考文献”字段按条进行换行处理,此任务也可通过替换完成。与上述类似,在元数据表中选中“参考文献”数据列,将2条文献中间的“.[”替换为“.
[”,导入网站后即可实现按条换行。“
”是浏览器默认换行符,具体替换内容可根据元数据格式适配。

打印从采编系统或网站导出的元数据时,一般要进行版式设置,常见的是页边距和页眉页脚设置。如图9所示,打开“打印预览”界面,勾选图10中的“页边距”,即可调整表格线位置,使内容打印在一张纸上。点击图10中的“页面设置”打开浮窗,选择其中的“页眉/页脚”标签,则可按需设置打印出的纸质文档中页眉页脚区域需显示的信息。亦可通过“自定义页眉”“自定义页脚”在左、中、右3栏分别显示由文档中提取的不同内容,兹不赘述。

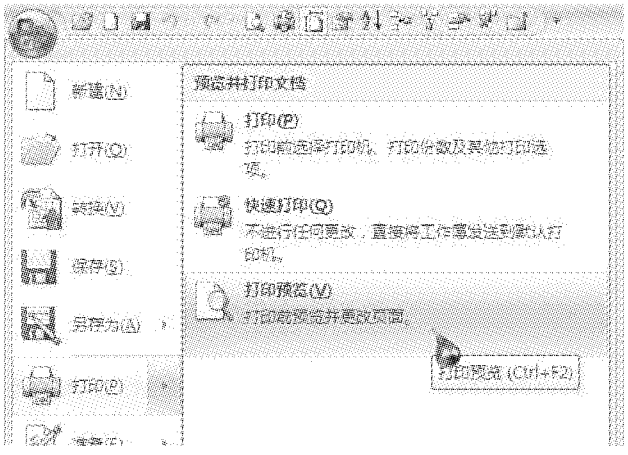

图9 打印预览选项位置

图10 打印页边距和页面设置

5 html元数据批量重命名

在网站上传html文件时,需按后台预置要求更改文件名方可正常识别,运用Excel的公式与拖动填充功能可快速生成用于批量重命名的.bat批处理文件。如图11所示,新建一个表并将前三列分别设置为“原文件名”“编号”“改名命令:复制到记事本,保存时扩展名改为.bat并放在html目录双击运行”。

图11 html批量改名文件设置

其中的“编号”数据由网刊发布系统中的“文章信息”复制而来,即网站要求的“年-期-起始页”html元数据命名规则。选中C2单元格,在上方“f_x”框中输入“="ren "&A2&" "&B2&".shtml"”,意为将第一列中的文件名改为第二列中给出的格式并将扩展名改为“.shtml”,如图12所示。

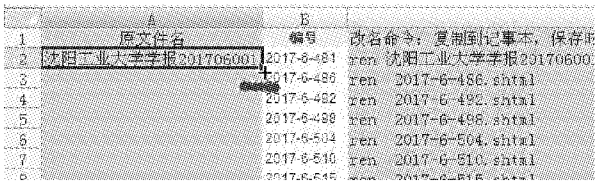

图12 利用公式生成重命名命令行

将原html文件名粘贴到第一行,并将鼠标停在单元格右下角变成黑色十字(见图12),按住左键向下拖动至所需位置,则“原文件名”一列会自动按照“末位数字+1.html”升序填充,同时C列会对应生成重命名命令行。同上文类似,按照C列标题提示的步骤选中所有ren命令行粘贴进“记事本”文件,将其保存为“html批量改名.bat”,放在原html文件夹中双击运行,即可将html元数据批量改为上传网站所需格式。(上述文件可至https://pan.baidu.com/s/

1t6YZb3fRVhYVVxmXOYu4Uw 下载后根据需要修改使用)

6　结束语

媒体融合的不断发展对编辑提出了更高的知识与能力要求,如何更好地运用现代化手段参与融合、提高工作效率值得深入探讨。更好地借助常用办公软件解决元数据处理问题,降低技术门槛和转换成本是今后值得持续关注的领域。

7　参考文献

[1] 陶艳, 董克. 基于计量的图书情报领域元数据研究现状分析[J]. 图书馆学刊, 2016, 38(4): 132

[2] 焦丽. 我国元数据研究述略[J]. 科技信息(科学教研), 2007, 24(35): 186

[3] 叶静. 从 2006-2011 年我国核心期刊载文分析看我国元数据研究新进展[J]. 科技情报开发与经济, 2012, 22(14): 126

[4] 赵秀君. 2000-2009 年我国元数据研究论文统计分析[J]. 情报科学, 2012, 30(2): 282

[5] 周亚. 2001-2008 年国内元数据自动抽取研究综述[J]. 科技情报开发与经济, 2009, 19(23): 140

[6] 汤敬谦, 杨鹤林. 热点、网络与态势: 国外图书情报学领域元数据研究的知识图谱分析[J]. 图书馆学研究, 2016, 37(6): 18

[7] 江燕青. 面向语义出版的学术期刊数字资源聚合研究[D]. 上海: 华东师范大学, 2016

[8] 董微, 赵捷. 开放期刊资源元数据质量管理研究[J]. 中国科技资源导刊, 2018, 50(3): 82

[9] 于倩倩, 张建勇, 黄永文. 大数据环境下的文献元数据标准设计特点分析[J]. 图书馆杂志, 2018, 37(11): 37

[10] 张志勇. 高校图书馆利用八爪鱼网络爬虫技术高效采集元数据[J]. 现代信息科技, 2019, 3(4): 4

[11] 崔玉洁, 廖坤. 借助八爪鱼采集器实现过刊网刊元数据的自动提取[J]. 编辑学报, 2016, 28(5): 485

[12] 黄政, 张学福. 一种基于网页信息抽取的 OA 期刊资源采集方法研究[J]. 数字图书馆论坛, 2017, 13(5): 25

[13] 杨海亮, 徐用吉. 提取方正排版文件广义元数据并生成全文 HTML 的探索[J]. 中国科技期刊研究, 2016, 27(2): 202

[14] 刘华中. 面向 PDF 文档的论文元数据提取方法研究[D]. 秦皇岛: 燕山大学, 2012

[15] 黄政. 开放获取期刊资源采集系统研究与实现[D]. 北京: 中国农业科学院,2017

[16] 冯民, 毛善锋. 一种适合大批量期刊元数据自动化提取的程序设计[J]. 中国科技期刊研究, 2016, 27(10): 1081

[17] 冯红娟, 李云龙, 梁蕙玮, 等. 面向资源检索的元数据仓储建设研究[J]. 图书馆学刊, 2015, 37(3): 34

[18] 赵悦. 数字图书馆元数据应用研究[D]. 武汉: 武汉大学, 2005

[19] 杨松迎, 王志鸿, 曹荣章. 科技期刊数字内容的挖掘与服务: 以《电力系统自动化》为例[J]. 中国科技期刊研究, 2017, 28(2): 145

[20] 贾欢. 科学数据元数据互操作研究[D]. 武汉: 武汉大学, 2017

[21] 伯琼. 元数据转换及集成的研究现状述评[J]. 青岛职业技术学院学报, 2007, 20(1): 54

[22] 陈春颖. 数据清洗技术在期刊元数据整合中的应用[J]. 图书情报知识, 2009, 26(6): 87

[23] 黄如花, 邱春艳. Dryad 数据仓储的元数据管理[J]. 图书馆杂志, 2014, 33(1): 68

[24] 吴丽杰. 电子连续性资源元数据标准建设研究[J]. 图书馆学刊,2013, 35(3): 35

[25] 吉海涛, 郭雨梅, 郭晓亮. 分学科高校哲学社会科学学术期刊数字化平台构建[J]. 出版科学, 2016, 24(6): 94

[26] 景勇, 郭雨梅, 丁岚. 科技期刊微信公众平台功能拓展[J]. 编辑学报, 2016, 27(4): 384

[27] 冯民, 高绍强. 用 excel 宏程序提取 fbd 期刊数据的简易编程[J]. 中国科技期刊研究, 2015, 26(9): 969

[28] 张志恒, 张显库, 杨光平, 等. 基于 Visual C + + 的 Excel 工作簿数据处理[J]. 软件导刊, 2017, 16(1): 135

(2019-06-03 收稿;2019-08-26 修回)