水电水利
Hydropower and Water Resources
第 4 卷◆第 2 期◆版本 1.0◆2020 年
文章类型：论文|刊号（ISSN）：2529-7821

水文年鉴数字化浅析

朱志强[1]　刘邑婷[2]
1 水利部海委漳卫南运河管理局　2 水利部海委漳卫南运河四女寺枢纽工程管理局
DOI:10.32629/hwr.v4i2.2742

[摘　要] 水文计算工作需要录入大量的文字、图像,然而水文年鉴均以纸质或电子书格式保存,使用人工录入投入大、错误率高,且使纸质水文年鉴逐渐磨损老化,因此,研究水文年鉴数字化对提高效率、加强纸质年鉴的保护和利用具有重要意义。本文提出了一种纸质及电子书格式转换为电子数据格式方法,应用结果表明,该方法识别精度较高,为水利工程的建设和管理、水资源管理和开发利用、水环境保护和水生态资源建设等工作提供了高效的服务。
[关键词] 水文年鉴；数字化；OCR

20世纪50年代,中国全面整编刊印了历史积存的水文资料,并将以后的资料逐年分区整理刊布。从1958年起,统一命名为《中华人民共和国水文年鉴》(以下简称《水文年鉴》),并按流域、水系统一编排成册。它集中反映当年的水文要素的变化过程,将数量庞大而又只有一份的原始水文观测记录,经过科学的分析整理编成简明的图表,节省了各个部门的统计加工,避免了大量重复工作。《水文年鉴》为水利工程的建设和管理、水资源管理和开发利用、水环境保护和水生态资源建设等提供不可或缺的基础资料,蓄水工程、调水工程、流域或区域的防洪规划、水资源综合规划等都是利用《水文年鉴》中的原始数据,进行流量、泥沙、降水等水文数据的分析,从而确定工程规模和规划方案[1]。20世纪80年代以前,《水文年鉴》只有刊印本,自2006年复刊以来,计算机技术开始应用于《水文年鉴》的储存和资料整理。肖卫等(2011)研究了特殊档案数字化加工系统,对水文年鉴进行数字化加工扫描并载入数据库,以推进水文年鉴电子信息化管理,实现水文年鉴资源依法共享[2]。

在水文资料印刷体文档中,表格是其必不可少的一部分,它简明规范地将所有的文档信息高度集中在一起,并且让读者准确地明白其表达的含义[3]。目前《水文年鉴》的表格形式均已固定,但由于其以纸质或电子书格式保存,不利于表格数据的摘录和使用。本文提出一种基于多种软件的《水文年鉴》数字化方法,最终实现表格数据的电子数据化存储。

实现数据电子化需要用到的软件有Apabi Reader、Apabi Maker、PDF Creator及ABBYY Fine Reader。

1 软件简介

1.1 Apabi Reader

Apabi Reader全称“方正Apabi Reader”,是一款面向电子书、电子公文、电子报纸、电子期刊等多种文档类型的阅读器。它支持CEBX/CEB电子图书的阅读、打印。

1.2 Apabi Maker

Apabi Maker是方正研究院数字出版分院开发的CEBX /CEB文件转换器。它提供了许多领域的电子出版服务,将包括传统的印刷图书在内的多种格式转成CEBX/CEB格式的电子图书供人们阅读。

1.3 PDF Creator

PDF Creator是一个在Microsoft Windows转换文件成PDF文件的软件。安装此软件后,用户选择 PDF Creator 作为虚拟打印机,允许大部分程序将文件转换成PDF。

1.4 ABBYY Fine Reader

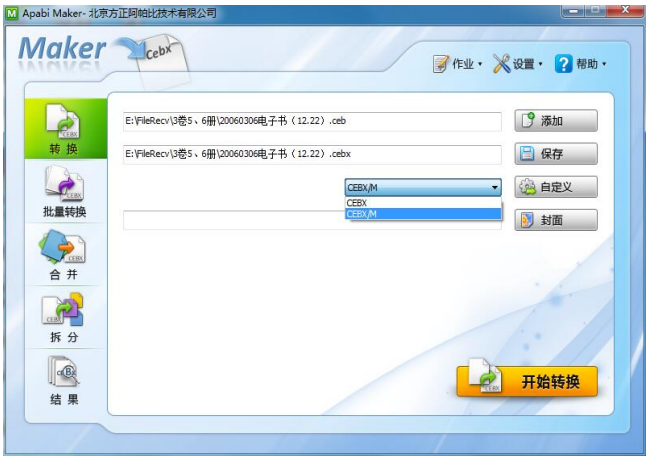
ABBYY Fine Reader是一款俄罗斯软件,在文档识别,数据捕获和语言技术的开发中居世界领先地位。其获奖产品 FineReader OCR 软件可以把静态纸文件和PDF文件转换成可管理的电子数据,可以大大节省您的时间和精力。2005年12月22日,ABBYY 美国宣布,ABBYY FineReader 8.0专业版被美国著名计算机杂志《PC Magazine》授予4星。ABBYY FineReader Professional 是一款真正的专业OCR,它不仅支持多国文字,还支持彩色文件识别、自动保留原稿插图和排版格式以及后台批处理识别功能,使用者再也不用在扫描软件、OCR、WORD、EXCEL之间换来换去了,处理文件会变的就像打开已经存档的文件一般便捷。由于《水文年鉴》文档中均以表格及数据形式存储,对OCR文本识别软件的要求较高。ABBYY Fine Reader作为一个OCR文字识别软件,该软件具有识别效率高、数据准确率高、表格一致性高的特点,完全满足对数据及格式的各项要求。

2 水文年鉴的数字化

2.1 CEB格式向CEBX格式的转换

目前《水文年鉴》全册电子版均以纸质或CEB电子书格式储存,CEB是Chinese E-paper Basic的缩写,是由北大方正公司独立开发的电子书格式,由于在文档转换过程中采用了“高保真”技术,可以使CEB格式的电子书最大限度地保持原来的样式。正是基于这种特点,国家有关部门将CEB格式作为电子公文传递的标准格式,许多电子书发行机构和数字化图书馆都已经开始采用这种格式。但是,CEB格式文件具有不能被编辑的局限性,实现水文年鉴的数字化,首先需要将CEB电子书格式转换为可编辑的CEBX电子书格式。

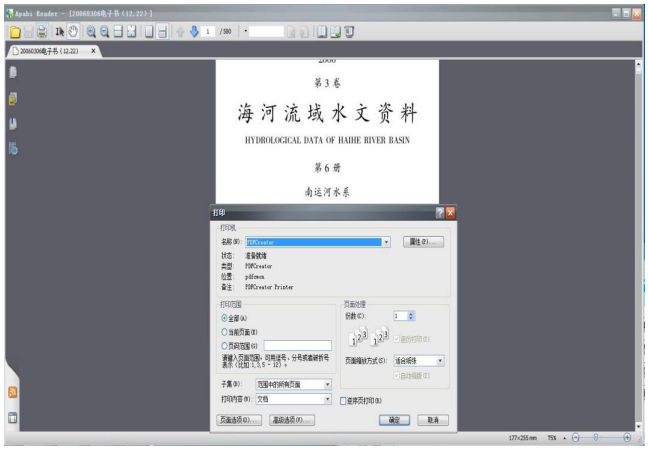
打开软件Apabi Maker,点击“添加”(添加需要转换的电子书)——“保存”(确定转换后文件存放路径)——“开始转换”。(注:转换格式缺省为“CEBX/M”,若出现转换失败的情况,点击下拉菜单选择“CEBX”格式继续转换即可。)



2.2 CEBX电子书格式向PDF格式转换

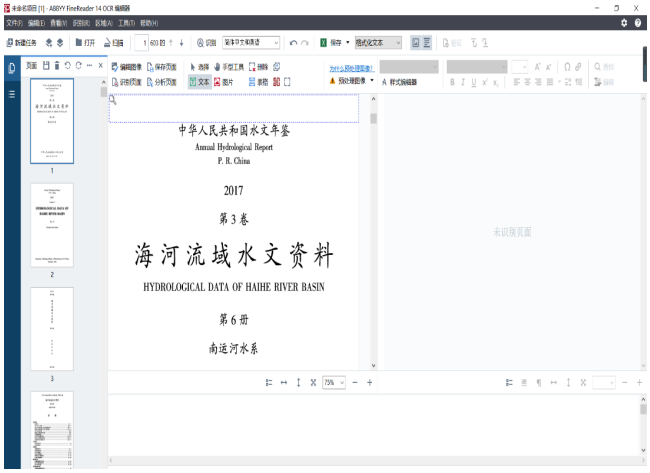
打开Apabi Reader软件,点击工具栏中的“打开”,找到刚转换成功的CEBX文档,之后点击“打印机图标”,打印机列表里选择PDF Creator,点击

“确定”,弹出PDFCREATOR对话框,选择目录后点击保存,便实现了CEBX电子书格式向PDF格式转换的目的。



2.3 PDF格式向XLS格式转换

打开ABBYY Fine Reader软件(本文以14版本为例),点击“工具”——“OCR编辑器”(或者直接在程序内打开“ABBYY Fine Reader 14 OCR编辑器”)。在OCR编辑器里打开“文件”——“打开图像”选择对应的PDF文件(选择需要转换的PDF文件,注:默认是将所有页面识别为电子数据格式,暂停识别点击“取消”即可)。ABBYY Fine Reader 14 OCR编辑器状态栏里有三列,左侧区域为缩略图视图,负责显示PDF文本的组织页面;中间部分为图像面板,负责显示所选PDF页面的内容;右侧部分为文本面板,负责显示识别成功的电子数据内容。



2.3.1建立模板

点击左侧缩略图内任一组织页面,选择工具栏内的“表格”,选择中间图像面板区域本页内容,将全部内容用表格框选起来。选择工具栏内“区域”——“保存区域模板”命名该模板(如:全表类)并保存。

2.3.2加载模板

点击左侧区缩略视图区域内任一组织页面,按键盘快捷键“Ctrl+A”,选择所有组织页面。选择工具栏内“区域”——“加载区域模板”,选择刚刚保存的模板(全表类)。这样便完成了全部组织页面的表格化,杜绝了文本当图片不识别或识别不完整情况的发生。

针对不同类型、格式的表格可以制作不同的模板,例如:全表格模板、逐日平均水位表模板、逐日平均流量表模板、逐日平均含沙量模板、逐日平均输沙率模板、逐日水面蒸发量模板、逐日水温表模板、逐日降水量模板及各时段最大降水量表模板。依据这些类别的模板可以达到事半功倍的效果。

2.3.3识别页面

点击工具栏内“识别”——“识别所有页面”,即可实现对PDF所有组织页面的识别。

2.3.4数据保存

点击状态栏内的“保存”选择保存Excel格式,这样就实现了《水文年鉴》文本的电子数据化。

3 结语

《水文年鉴》为纸质格式时,可将文本扫描为PDF或JPG格式保存,使用ABBYY Fine Reader同样可实现其数字化。

《水文年鉴》的数字化既减少了对纸质年鉴的损害,保护了珍贵的历史水文资料,又实现了表格数据的提取和存储,提高了工作效率。应用结果表明,本文所述方法具有较高的识别精度,方便水文计算中数据的调用,为水利工程的建设和管理、水资源管理和开发利用、水环境保护和水生态资源建设等工作提供了高效的服务。

[参考文献]

[1]彭立波,鲁青.特殊的水利档案信息资源——水文年鉴[J].黑龙江水利科技,2009,(6):135-136.

[2]肖卫,时昶.水文年鉴数字化研究[J].水生态学杂志,2011,32(4):149-151.

[3]陈婉婉,李士进,胡金龙,等.一种新的水文年鉴数字化方法[J].计算机与现代化,2016,(12):78-82.

作者简介:

朱志强(1985-)男,山东省东营人,汉族,本科,工程师,主要从事防汛抗旱和水文预报工作。