

· 杂谈 ·

高校纸质档案光学字符识别与数字复制研究

□ 李 莹

摘 要:传统的高校档案以纸质为主,为紧跟时代发展、提高工作效率,高校不断改进传统的纸质档案复制方法,信息化、数字化电子档案逐渐完善。利用光学字符识别技术对扫描保存的图像进行处理,修复破损模糊问题,并进行文字信息的搜索识别,用数字复制档案查询利用方式减少对历史纸质档案原件的查阅使用率,保护珍贵历史,重视档案管理。

关键词:高校;纸质档案;光学字符识别;数字复制

传统档案以纸质档案为主,由于早期归档不科学,时间久远导致纸质风化、泛黄、模糊、破损,再加上库房档案数量冗多,借阅查询时费时费力,影响档案管理与服务利用效率。区块链、5G 网络等信息化技术的发展,加快了纸质档案进行数字复制的进程,也有效提高了档案管理人员的专业水平和查阅效率。

国家档案局 2019 年末发布了《纸质档案数字复制件光学字符识别(OCR)工作规范》,规范了对纸质档案数字复制工作的具体实施流程,使用科学手段最大化纠正手写纸质档案存在的若干问题。纸质档案扫描整理要按照一定要求,对已经破损、模糊的档案一定要及时进行修复,并制定专业的指导方案。在工作开展过程中,计算机专业技术人员和专业档案管理人员需要共同协作完成数字扫描、图像处理、字体认证、目录著录、原文挂接、数据上传及备份等工作,注重档案信息检索和数据更新利用的原则。此外,不可或缺的是对档案从业较久的人员进行数字化过程监控,如根据档案实际对内容进行补充著录、题名关键词的提取、检索等工作。

1 光学字符识别技术及纸质档案数据复制工作

光学字符识别(Optical Character Recognition,简称 OCR)是指通过识别算法来分析图像中的有效数据,可对机打、手写字体进行识别,扫描图像后自动生成文字数据,可被计算机识别检索的技术。档案光学字符识别扫描工作,应对纸质档案进行整理定位,数字复制的电子文件应便于系统挂接查找。涉密档案应按照相关管理规定和技术要求,规范开展光学字符识别与数字复制工作,加强过程监管、质量管理和安全保障,建立紧急问题预警与处理机制,以档案信息安全为首任,确保数字复制成果质量。

档案光学字符识别扫描复制不是一项孤立的工作,应当做好顶层设计,做好制度规范、资金预算、人员配备、技术指导、过程监管,统筹规划、有序实施,逐步实现数字档案的常态化。具体工作计划应从图像处理、

信息识别、修复认证、输出保存等流程进行探索。

第一步是图像处理,主要是把扫描后的图像按照识别的要求进行格式、大小等修改,并在扫描完成后对部分原本倾斜的纸质档案进行图片矫正。通过图片降噪处理、灰度图片设置等最大程度还原纸质档案扫描图像中的有效信息。这项基础工作是光学字符识别的第一步,也是关键环节,要求扫描生成图像辨识度和清晰度达到一定标准,符合设备的基本设置要求,图像处理标准与否决定了后续数字化工作的效率。

第二步是信息识别,主要是对纸质档案文字、图像等信息进行校对、纠错,以满足关键词、查询内容的信息快速识别。信息识别是查询工作的重要环节,可以分为文字识别、图像识别、表格识别、印章识别等。其中,由于档案记录以文字为主,后期检索也是通过关键词进行查找,所以文字识别工作占主体地位,文字包括中文、数字、英文和字符的识别。数据词库范围、专业、固定搭配以及定期更新词库等,是保障和提高识别查询速度、效率及维护的具体工作。

第三步为修复认证,主要针对信息识别阶段无法识别检索的文字、图片等信息。由于手写纸质档案字体多样,部分信息难以辨认,加入修复认证环节,通过光学识别软件进行智能化的字符信息转化识别,提供相似结果的信息筛选,便于工作人员经过辨认较快选择正确信息,减少扫描过程的机械错误。在保存筛选数据结果的同时,要做好原扫描图像的存储备份,如果后期检索过程中出现不确定结果时,需要再次对原图进行验证。

第四步为输出保存,将扫描的电子文件输出为两种格式进行保存,一类是原图像扫描件,另一类是光学字符识别的纯文本格式。后期使用以纯文本格式为主,因为文字信息的词库可以打造强大的数据库信息,使数据词库之间建立有效联系,档案数据还可以协同其他职能部门的信息数据,使原本孤立的扫描件信息融合成为一个整体数据库,转化输出为更有价值的新数

据,满足部门之间的信息输出共享,为数字高校的建设奠定数据基础。

2 急需进行光学字符识别的纸质档案类别

受存放年代时间久远、纸质载体容易损坏和档案保存条件标准不高等因素的影响,目前很多高校档案馆保存的历史档案出现了字迹模糊不清、纸张损伤严重、部分污浊浸染等状况。因此,高校传统纸质档案的数字化工作显得尤为紧迫和重要。其中查阅率较高、年代较久远的主要有学生的学籍档案、综合文书档案、基建档案及声像档案四大类,下面将围绕这四类档案进行具体阐述,将传统纸质档案通过数字扫描方法转化为虚拟的电子文件形式,实现档案管理系统平台的数字化利用与储存,解决当前高校历史纸质档案的存储危机,为档案管理者 and 利用者查阅档案提供便利。

2.1 学籍档案

学籍档案主要包含学生的学籍卡和成绩册。按照时间顺序主要分为两类,时间较久的为一类,以手写体为主,由于书写笔迹和纸张老化等原因,20世纪80年代的学籍信息存在字迹模糊、学籍照片受潮变花或粘在一起、没有成绩记录、教务系统查询不到信息等问题;另一类则是电脑普及之后打印版本的学籍档案,从教务系统里下载PDF版本再打印盖章保存,电子版也会同步上传至档案系统,但是电子版缺少教务处的章,并不具备法律效力,没有起到真正的数字化效果。所以,要通过技术手段对盖有红章的纸质学籍卡和成绩册进行光学字符识别扫描,修改、校正模糊的文字和照片后再上传至档案系统,方便利用。

2.2 文书档案

文书档案包含高校各部门的综合业务档案,从建校的批复文件到人才评聘再到各类会议纪要,记载了学校的发展历程和重要事件。尽管有些档案已经存放了几十年,但是后期学校发展升格、人员退休等还会使用到这些重要信息,不能因保护不善而影响利用。纸质档案会风化变脆,不能及时数字化扫描处理势必会出现部分档案破损、模糊现象,急需管理者予以重视,积极响应国家政策,落实开展纸质档案的数字化工作。根据《档案服务外包工作规范》聘请专业的外包公司进行合作,并由资格较老的校内人员对档案材料进行鉴定,档案人员配合外包公司做好档案安全、整理、装订等工作。很多文书档案全文信息较多,仅仅通过图片查找其中一项的话较为困难,如录取名册一年有几千名新生名册,肉眼识别很容易漏掉,通过光学字符识别的文字识别功能可以实现快速精准查找。

2.3 基建档案

基建档案是记载学校建筑建设的重要依据,包含各项工程审核材料及图纸信息,在改、扩建、维修等情况都需要查阅相关材料。图纸类档案近几年才有配套的电子版,早期的蓝图只有纸质版,而且尺寸较大需折叠存放,折叠边角经过反复查阅容易磨损,蓝图时间久了也会出现风化、泛黄、变脆等老化现象。通过光学扫描对纸质图纸等档案进行还原识别,确保档案在继续使用的同时避免直接接触纸质原件,起到保护作用。

2.4 声像档案

声像档案主要包含照片、光盘等。老照片承载了一所学校的记忆,在校庆活动上经常会展现,但是老照片存在受潮、变花、粘连等问题,如不及时修复,将很难修复利用。而档案管理人员的拍照和图片修复技术不甚专业,无法满足数字化扫描并存储利用的要求,需要聘请专业人事负责开展工作。光盘作为电子图片和视频的主要载体,易于刻录和保存,但是需定期进行检查和复刻,以保证信息的完整。

3 结语

光学识别技术在档案管理方面起着承前启后的作用,及时解决老旧纸质档案的消损带来的负面影响,为纸质档案数字复制工作提供了有效的技术手段和正确的信息匹配,满足电子档案的存储与快速查询利用,协同了各部门之间的数据链接,促进了数字高校建设的整体进度与效果。其在对纸质档案的保护与传承中发挥着巨大作用,同时提高了档案管理人员开展工作的效率,希望高校档案部门能够认真策划、组织学习,切实落实纸质档案的数字扫描复制与查询利用工作。

基金项目:2020年度江苏高校哲学社会科学研究项目《5G时代受众理论视域下的高校档案服务创新发展研究》(项目编号:2020SJA1125)。

参考文献

- [1]丁德胜.《纸质档案数字复制件光学字符识别(OCR)工作规范》解读[J].中国档案,2020(02):40.
- [2]赵宗明.高校学生档案信息的功能及开发利用[J].产业与科技论坛,2020(14):219.
- [3]闫茹,孙永奇,朱卫国,等.基于CNN与有限状态自动机的手写体大写金额识别[J/OL].计算机工程:1-12[2021-06-03].<https://doi.org/10.19678/j.issn.1000-3428.59118>.

(作者单位:徐州工业职业技术学院图文中心)