

古籍记录历史、传承文化,是中华文明源远流长、博大精深的表征和见证。随着古籍数字化步伐的加快,实现全部古籍永久保存的目标有望实现。越来越多收藏在图书馆里的珍贵古籍走出“象牙塔”,走进社会大众。古籍数字化的持续推进,让古籍面貌焕然一新,不断激发古籍生命力。“活起来”的古籍日益成为传承中华优秀传统文化、坚定文化自信的宝贵滋养。

古籍数字化带来阅读便利

什么是“天头地脚”?什么是“象鼻”?什么是“鱼尾”?……打开识典古籍网站《永乐大典》数字高清影像库,点击这些看起来有些陌生的名词,《永乐大典》高清图片上相应的位置立刻突出显示,直观而形象地展现这部珍贵古籍的风采。更令读者感兴趣的是,网页上的《永乐大典》可以随着鼠标调整方向和角度,360度观赏,仿佛是拿着真书在阅读。

“这是我们的专利技术,用光影变化最大限度地模拟实体书的阅读体验,最大限度地保存古籍原貌。”识典古籍项目产品负责人王宇说。

《永乐大典》数字高清影像库是国家图书馆承接的国家古籍数字化重点项目,由北京大学数字人文中心和字节跳动公司共同设计研发。该项目第一辑收录国家图书馆馆藏《永乐大典》40册、75卷的内容,除呈现《永乐大典》高清图像、整体风貌及相关知识外,还尝试对部分大典内容做了知识标引,为后续《永乐大典》的知识体系化、利用智能化进行探索。

“数字化解决了存藏和使用之间的矛盾。”国家图书馆副馆长、国家古籍保护中心副主任张志清说。因古籍年代久远,极易破损,珍贵古籍的借阅有一整套严格的流程,普通读者是难以接近的。“纸寿千年”,每一次翻阅都是对古籍的伤害。“古籍数字化既减少了纸书的磨损,也使《永乐大典》这部宝贵文献‘化身千百’,走入寻常百姓家”,从而弘扬中华优秀传统文化,推动相关学术研究,感知《永乐大典》的不朽神韵。”张志清说。

除了让珍贵古籍走近普通读者,数字化也降低了古籍的阅读门槛。

在识典古籍网站和客户端上,《论语》《孟子》等常见古籍不仅有文本,而且有古籍原本影像,图文左右对照;有注疏和翻译,文白对照,不仅可以阅读,还可以检索。“目前我们初步把人名、地名和官职等实体进行了标注,也上线了字典释义功能,鼠标或手指放到不认识的字词上,就会显示注释。下一步还要把实体标注与百科词条链接起来,遇到不懂的问题,点击即可呈现详细解释。”王宇说,通过数字化降低古籍阅读门槛,让更多读者亲近古籍,感受中华优秀传统文化的魅力。

作为面向大众的公益性古籍数字化平台,目前识典古籍已上线古籍1600余部,免费对公众开放。读者对于数字化的古籍表现出很高的阅读热情,短短数月,识典古籍累计用户数已超过1240万。

面向专业用户的古籍数据库建设步伐也在加快。中华书局下属古籍数字化企业古联公司开发的古籍整理出版资源平台籍合网,自2018年上线以来,已发布数据库31个,涵盖专业古籍整理出版资源20亿字,石刻资源5万余篇,历代登科人物10万余条,木版年画18000余幅,书法作品10000余种,甲骨文卜辞143856条,总计字符30多亿。

自2012年以来,我国古籍事业进入新时代,古籍数字化不断提速。2022年4月,中共中央办公厅、国务院办公厅印发《关于推进新时代古籍工作的意见》,明确提出“推进古籍数字化”,强调“支持

古籍数字化重点单位做强做优,加强古籍数字化资源管理和开放共享。”《意见》进一步激发了相关单位古籍数字化的动力,以大型图书馆为主体的公藏单位普遍开展了古籍数字化工作。

今年9月,国家图书馆组织第八次古籍数字资源联合发布,新增发布古籍资源1672部(件)。至此,全国累计发布古籍及特藏文献影像资源达13万余部(件)。

不过,在快速发展的同时,古籍数字化还存在薄弱之处。据统计,我国现存古籍约20万种5000多万册(件),但实现数字化的不超过8万种,大多数所谓数字化古籍只是完成了初步的影像扫描,真正实现文本数字化的不足4万种。

古籍数字化专家吕亚峰说,古籍影像扫描是古籍数字化的基础,但影像无法检索;只有数字化文本才可以检索,便于研究、阅读和传播。因此,今后古籍数字化工作应把重点放在古籍文本的数字化上。“现在有了人工智能等先进技术,加速实现全部古籍的数字化已经具备了条件。”吕亚峰说。

“如果现存古籍全部数字化,那么我国古籍保护就迈上了一个新台阶,古籍灭失的可能性就大幅降低了。这对于赓续中华文脉,将是一个了不起的贡献。”张志清说,“我们所处的新时代,有史以来,中华民族第一次有可能实现文化典籍永久保护和传承。”

古籍数字化进入人工智能时代

将古籍高清图像上传服务器,点击自动识别按钮,只见古籍影像上立刻出现一个个不断闪动的格子,自动套住图片上的每个字,相应的文字就按照古籍上文字的排列顺序出现在页面。这一过程不过数秒。

“人工智能在古籍数字化中的应用有效提升了古籍整理的效率。”古联公司总经理洪涛说,以籍合网OCR(光学字符识别技术)识别为例,5分钟的OCR识别相当于人工录入20小时的工作量,而且错误可以降低75%。“这对传统纸本古籍整理效率的提升是巨大的,”洪涛说,人工智能解决了大型古籍整理项目耗时耗力、过久过多的现状。

以西泠印社“刻在石头上的浙江”系列丛书为例,古联公司编辑部参与了包括《东瓯金石志》等在内的9种金石类古籍的整理,共计约147万字。在整理过程中,全流程利用智能整理技术辅助工作。首先用OCR识别底本文字,然后采用线上众包模式开展底本校对工作。校对后的稿件通过自动标点功能进行标点,标点后的稿件再交由编辑进行校勘和通读审稿。在这种工作模式下,编辑部仅用时4个月就完成了全部稿件的整理工作和部分编辑工作。“这样的速度在以前是不可想象的,只有在人工智能时代才能成为现实。”洪涛说。

人工智能正在变得越来越“聪明”,现在对版刻本的识别准确率可达98%。即使对行夹注、眉批、行间批注、表格等不规则的古籍版面,不仅能准确识别文字,还能通过针对性训练优化分区效果,从而避免出现虽然单字识别正确,但阅读顺序颠倒错乱不能复用的情况。

以往给古籍断句和加标点,需要经验丰富的专业人士手动进行。但现在人工智能可以实现机器自动断句、自动标点。洪涛介绍,目前研发的自动标点技术,断句准确率平均达到98.46%,标标准准确率93.94%,专名线书名线自动标注准确率达到92.15%。标点标线后的古籍文本更方便阅读、理解和研究。

“如果人工智能不介入,完全依靠人力,古籍数字化全部完成可能还需要上百年的时间,但有了人工智能,完成全部古籍的数字化也许只要二三十年。我们对此有信心。”王宇举例说,百衲本《二十四史》将近4000万字,但利用人工智能,识典古籍北大整理团队仅用3个多月就完成了识别、点校、上线发布。

尽管经过人工智能处理的古籍文本准确率已经很高,但依然需要人工核校。特别是一些异体字、冷僻字,以及特殊格式的古籍,更是离不开专业古籍工作者的核校。

“人工智能并不能完全取代人,人工智能的优势在速度,而人的优势在精确和创造性。通过人机协同、人机互补,可以大幅提高古籍整理出版的效率和质量。而高质量的、学术性的古籍整理工作依然需要专家学者长期不懈的努力,不可能被计算机所替代。”洪涛说。

古籍整理进入大众化时代

古籍整理是专业性很强的工作,但在人工智能高速发展的今天,古籍爱好者也可以参与古籍整理。其流程大致是这样的——

经过人工智能处理的古籍文本,通过互联网分发给在线编校人员,后者在电脑上对照古籍底本高清图像,逐字逐句审阅,并在古籍整理平台上修改。管理员借助平台的版本比对功能审阅修改记录,判断是否合格。

古联公司建立的全国第一个古籍整理众包平台正是这样运作的。从2018年上线以来,籍合网古籍整理众包平台积累了近5000人的在线古籍编校队伍。这些在线古籍编校者绝大多数都是业余人士,有大学生、图书编辑、律师、教师,也有公务员、金融界人士、自由职业者等。爱好文史、熟悉互联网是他们的共同特征。

全书近2亿字的《中华大藏经·续编》通过众包模式,在全球900多位业余审核者的参与下,仅用时两年就完成了两个校次的底本校对与审核工作。“古籍整理众包工作模式不仅解决了大型古籍整理项目在人力方面的需求,更重要的是解决了地域限制和个体时间分散、随意的难题,使人力和时间都得到了充分的利用。”洪涛说。

“青年一代对中华传统文化的认识和热爱,是推动古籍事业高质量发展的重要推力。”张志清说。

目前,一支2300余人的古籍志愿者队伍活跃在字节跳动公司。每次识典古籍发布古籍审核任务,志愿者们就会在群里认领,往往是“秒光”。手速稍慢,任务就被人领走了。

古籍志愿者许霞丹参与了《茶苑》和《鸡足山志》两部古籍的文字校对和结构整理任务。“亲自参与,大受震撼,完全颠覆了我的认知。原来,只是轻轻滑动鼠标就能浏览到的内容,背后凝结着这么多复杂的步骤和这么多人的心血。”许霞丹说,“除了为古籍数字化做出自己的微薄贡献,还结交了志同道合的朋友,自己也增长了不少见识。我们都在期待下一次的古籍整理任务了。”

在人工智能的帮助下,未来的古籍数字化还会有怎样的进展?

洪涛认为,目前的古籍数字化仅仅是一个开始,未来可以利用人工智能进行古籍的辑佚、汇编、注解、翻译、检索、索引甚至考证和摘要工作。而更重要的是,利用人工智能开发古籍,用古籍蕴含的浩瀚素材创作生成内容。

“比如,创作一部历史题材的影视剧就可以让人工智能从古籍里寻找相关素材,自动生成故事。创作者在此基础上继续修改。这个过程不断循环,就能创作出高质量的作品。我们离这一天已经越来越近了。从这个意义上说,中华古籍是一座宝藏。”洪涛说。

数字化,激活古籍生命力

本报记者 张贺



活化利用 古籍重光

曾宇昕

古籍是中华文明的瑰宝,是中华民族的精神财富。习近平总书记指出,“深入挖掘古籍蕴含的哲学思想、人文精神、价值理念、道德规范,推动中华优秀传统文化创造性转化、创新性发展。”做好新时代古籍工作,对赓续中华文脉、弘扬民族精神、增强国家文化软实力、建设社会主义文化强国具有重要意义。

古籍蕴含古文字奥义,加之排版、文字、语法不同于今天的白话文,对于普通大众而言,存在一定的阅读门槛和理解难度。因此,推进古籍活化,成为新时代古籍工作的重中之重。《关于推进新时代古籍工作的意见》明确,要完善古籍工作体系,提升古籍工作质量。古籍活化让古籍走出学术“象牙塔”,走入千家万户,实现古籍服务大众、服务社会的目标。

推动古籍活化利用,数字化技术的力量不可或缺。随着数字技术的发展,许多图书馆、博物馆、出版社和古籍收藏机构等依托馆藏资

源,通过建设数字化平台、举办展览、开展竞赛、开设课程讲座等多种方式活化古籍资源。不久前,国家图书馆(国家古籍保护中心)组织联合发布了第八次古籍数字资源。古籍数字资源的在线发布与共建共享,让广大专业研究者与历史文化爱好者可以足不出户浏览古籍资料,推动古籍资源的深度开发、利用与传播。

在新媒体主导的“图像时代”,中华优秀传统文化的传承和发展正经历一场视觉转向。《典籍里的中国》《古书复活记》《穿越时空的古籍》等文化类节目的创作与播出,体现了用影像激活文化典籍,以当代受众喜闻乐见的形式传承和传播传统文化的有益探索。这些蕴藏在中华古籍中丰富而深邃的思想,以影像创作的方式提炼、呈现出来,对人们增进民族文化认同、提升民族文化自信、塑造正确的人生观和价值观具有积极作用。

类型多样、形式活泼的古籍活化方式,让古老的文字文本呈现出

新颖且充满趣味的面貌。国家图书馆推出“妙手匠心——修复互动体验套装”系列文创,让大众在修复体验中学习除尘、补缺、溜口、折页等古籍修复步骤,引发对古籍修复事业的关注。

此外,还有一些传统文化爱好者以古籍记载为依据,复原传统服饰、再现古书中的餐饮、利用古籍中的字体打造艺术字。古籍中精深微妙的思想,只有人们亲身阅读、走入古籍,方能领略其中的奥妙和真意。因此,多样化的古籍创新表达方式是“引”,更重要的还是由此引发大众广泛的阅读兴趣,引导人们真正投入古籍文本的阅读活动中。

卷帙浩繁的中华文化典籍,犹如中华文明的源头活水,百川汇聚、奔涌不息。进一步推动古籍的活化利用工作,让书写在古籍里的文字活起来,推动中华优秀传统文化创造性转化、创新性发展,必将为实现中华民族伟大复兴提供强大精神力量。

图①:浙江湖州嘉业堂藏书楼俯瞰。

图②:山西第一期古籍修复培训课堂上,学员听专家讲解相关知识。

图③:《永乐大典》数字高清影像库项目工作人员进行古籍数字化处理。

图④:《永乐大典》数字高清影像库。

图⑤:《永乐大典》数字高清影像库。

图⑥:《永乐大典》数字高清影像库。

图⑦:《永乐大典》数字高清影像库。

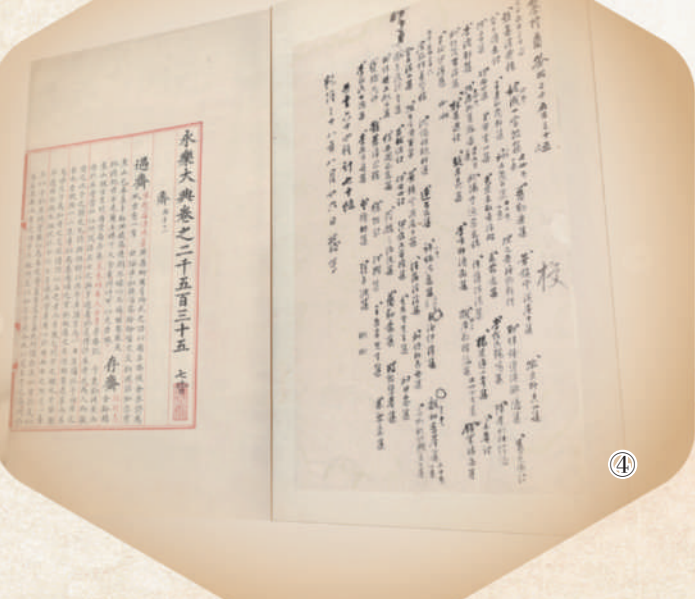
图⑧:《永乐大典》数字高清影像库。

图⑨:《永乐大典》数字高清影像库。

图⑩:《永乐大典》数字高清影像库。

图⑪:《永乐大典》数字高清影像库。

图⑫:《永乐大典》数字高清影像库。



图④

图⑤

图⑥

图⑦

图⑧

图⑨

图⑩

图⑪

图⑫

本版设计:汪哲平