·文献学的未来笔谈·

古籍资源的数字化与智能化开发利用

北京大学数字人文研究中心 王 军

古代典籍是中华文明最重要的承载体。今天我们说中华文明是古代文明唯一没有中断的文明体,就是因为中国古代典籍虽然历经焚毁、战乱、掠盗、散佚等等劫难,但是在历代文化人精心守护之下得以延续至今。金石甲骨、绢帛简牍、抄本刻本,每一次文献载体的演变不仅关系到典籍的命脉延续,而且对知识生产、文化传播乃至教育普及都影响甚巨。

在这样的大背景下来理解古籍数字化,它就不再仅仅是一个数字化技术解决方案,而是当代人将传承数千年的典籍文献迁移到数字信息环境下的历史责任。与以往任何静态的文献载体不同,计算机技术支撑下的数字信息环境不仅是信息的承载体,它同时兼具信息处理、分析、交互、传播等全功能,它本身还是能动的,而且在快速地自我进化。也就是说,与以往各类静态的信息载体不同,今天的数字化信息环境不仅造成纸本文献向数字形态转移,同时也引发文献的处理方式、操作方式、传播方式的快速演化。特别是愈加成熟和普及的人工智能技术,将造成人类知识生产和传播的革命性变革。相应的,文献学的研究手段也将随之改变。

基于对数字信息环境的这种理解,北京大学成立了“数字人文研究中心”这一校级虚体研究机构,为人文学科、社会学科和信息技术学科的研究人员提供跨学科协作平台,目标是打造数字环境下支持人文研究的数字人文基础设施。在古籍数字化领域,我们在古籍智能处理技术、历代目录集成、古籍数字化产品设计、古籍知识图谱重构、古籍大规模阅读平台建设等多个层面进行布局,为古籍向数字环境的迁移提供全面解决方案。

在古籍智能化技术方面,我们利用深度学习技术在数十亿字的古代标记语料上迭代训练语言模型,实现对古代文本的自动标点、专名识别和自动分词。我们构建的“吾与点”古文智能处理系统(http://wyd.pkudh.xyz)向大众提供开放服务。在先秦到明清的各类文本上,上述各项功能的平均准

确率达到95%左右，展示了智能技术在古籍整理领域的巨大应用潜力。

在古籍目录集成方面，我们构建了“历代史志目录集成分析系统”（http://bib.pkudh.org），对八种官修目录所包含的全部古籍书目记录做结构化、规范化处理和书名认同，在此基础上对历代官修目录的类目演化和书目流传做可视化展现，并提供各类检索和统计功能以便利研究人员在大规模古籍目录数据上展开研究，以数字手段辅助实现目录学“辨章学术、考镜源流”的学术功能。该项目的数据集包括：从《汉书艺文志》到《清史稿艺文志》的七朝正史艺文志，加上《四库全书总目》和《中国古籍总目》，贯穿了汉朝以来的主要历史朝代，数据总量达50多万条。

在古籍数字化产品开发方面，我们承担了国家图书馆和国图出版社的“《国家珍贵古籍名录》知识库”（http://rarebib.pkudh.org）和“《永乐大典》高清影像数据库”（http://yongle.shidianguji.com）的设计和开发工作。前者将珍贵古籍名录书目数据重构为知识库，综合应用了时空分布、多维检索、关系发现等多种数字人文技术，通过在历史时空中连续展现存世珍贵古籍版本的流衍过程，将古籍资源转化为文明演化的历史证据。后者在四十册《永乐大典》高清图文对照阅读的基础上，用文字、声音、动画集成地展现大典书册的生成、府藏、流传、散佚和收集的全过程，在交互设计和界面设计上做到美轮美奂，是古籍数字化产品的典范。

在古籍知识图谱重构方面，我们将240万字的宋代理学巨著《宋元学案》全面改造成知识图谱，将学案记载的二千馀学人及其间关系，以及他们与时间、地点、著作间的复杂语义联系提取出来编织到图谱中，将皇皇巨著以可视化、交互式浏览、语义化查询的方式提供利用。“《宋元学案》知识图谱可视化系统”（http://syxa.pkudh.org）一方面为学者提供了知识化、语义化的分析和研究工具，便于学术研究的开展；另一方面为读者打造了一种全新的阅读体验，为互联网一代接触、了解、学习宋代学术思想提供了路径。

在古籍资源库建设方面，我们与字节跳动合作，打造互联网环境下面向大众的大规模古籍阅读平台——“识典古籍”（http://www.shidianguji.com）。初步计划在三年内实现一万种基本核心典籍的数字化整理，在平台上向全网提供公益性的开放访问。不同于现有的各类古籍图文库，识典古籍平台希望能将古籍的阅读体验提升到现有互联网流行产品的水平，使得古籍资源的查询、阅读和利用符合互联网一代的使用习惯。

以上各领域的尝试，分别带来古籍整理、目录学、版本学、内容挖掘等文献学分支学科的技术性转变，提示未来在数字化技术支撑下文献学可能迎来新的发展方向。

我们认为：古籍的文本化和全文检索仅仅是古籍数字化的第一步。古

籍数字化的终极目标应该是将古籍资源改造成适宜在数字信息环境下保存、传播和利用的知识资源，也就是说，要为数字环境下古籍资源的生产、管理、利用、传播等提供全流程的系统解决方案。具体包括：

（一）充分利用前沿技术，特别是人工智能技术，加速古籍的数字化进程。在文字识别环节，OCR 技术在刻本文献上已经达到实用化的程度，目前的工作重点是抄本手写文字识别以及复杂版面切分；在自动句读/标点方面，通过人机协作的方式可以有效提升古籍标点的效率；目前只有少数古籍出版社在整理本中标记专名，随着古文命名实体识别技术的进一步提升和普及，古籍中人名、地名、职官、书名等专名标记会成为古籍整理的标配。

（二）现有的古籍整理工作主要是面向古籍整理本出版的，现有的古籍数据库以满足古籍查阅和学术研究为主要目标。**要在互联网环境下更好地传播古籍文化，需要以互联网产品的思维来设计和运营古籍数字化作品。**根据互联网用户的需求和行为习惯提升古籍数字化作品的用户体验，凭借融媒体的特性打造强交互的、生动的古籍数字产品，并通过互联网展开口碑传播。

（三）智能技术的普遍应用将大大降低古籍整理的专业门槛和成本投入，古籍数字化和古籍数据库建设将出现多种生态。以往主要依靠专业的古籍出版社和古籍数据库商来完成古籍整理出版的工作，投入大、周期长，加上古籍整理人才严重不足，导致古籍整理工作进展缓慢。古籍智能技术将使得各类古籍收藏机构、互联网企业，甚至古籍爱好者也加入到古籍数字化和古籍整理的事业中来。这不仅将加快古籍数字化进程，也会使古籍数字化产品的谱系更加丰富。

（四）古籍知识图谱自动构建、文言和白话的自动转译、基于古代语料的图文问答系统，将是古籍数字化领域下一步要突破的技术难点。这些关键技术的成熟运用，将使得大批古籍资源以关联数据的形态全面融入到互联网信息资源中，从而给数字时代的古籍资源开发与利用带来两方面的深刻影响：（1）古籍资源的形式将不再是单一的古籍数据库或整理本文献，而会以文化历史知识问答、名言名句引用、碑文石刻解读等各类互联网产品的形态出现，真正实现古籍的活化应用；（2）古籍知识图谱数据的融入将使得互联网中文信息资源达到前所未有的知识深度和关联广度，为人工智能算法提供丰富的历史文化语料，从而大大促进历史文化领域人工智能的应用水平。这是古籍数字化对于提升整个互联网信息环境的价值。