世界互联网大会 World Internet Conference

中国人民大学 RENMIN UNIVERSITY OF CHINA

世界互联网大会文化遗产数字化专委会系列研究

全球视野下的中国古籍数智化演进与未来趋势报告

中国人民大学数字人文研究院

2026 年 4 月


世界互联网大会
World Internet
Conference


中國人民大學
RENMIN UNIVERSITY OF CHINA

中国人民大学工作组

组长

夏翠娟　中国人民大学信息资源管理学院教授

工作组成员

龙家庆　中国人民大学信息资源管理学院讲师
严承希　中国人民大学信息资源管理学院讲师
胡云怡　中国人民大学信息资源管理学院讲师
郑振魏　中国人民大学信息资源管理学院博士生
曲梓萌　中国人民大学信息资源管理学院博士生
孟令国　中国人民大学信息资源管理学院博士生
居思微　中国人民大学文学院博士生
吴世强　中国人民大学文学院硕士生
祝章霞　中国人民大学文学院硕士生

作者贡献

夏翠娟：调研团队组织、报告内容框架搭建、引言和结论撰写、全文统稿修改审定。

龙家庆、曲梓萌、郑振魏：全球视野下中国古籍数字化存量与增量。

严承希、郑振魏、孟令国：技术驱动的中国古籍数智化开发路径和利用模式。

胡云怡、孟令国、曲梓萌：国际合作和人智协同标准的制订和应用。

居思微、吴世强、祝章霞：“中国古籍”定义、调研范围框定。

目录

一 引言

在 AI 技术高速发展和向垂直领域全方位渗透的背景下，中国古籍的保护与传承已步入从“数字化”向“数智化”跨越的关键窗口期。中国人民大学数字人文研究院通过对全球范围内多个国家和大型文化记忆机构及科研机构的深度调研，编制本报告。报告旨在通过详实的数据调研与案例分析，全面审视中国古籍在数字空间的存在形态、技术范式及传播效能，为构建全球性的中国古籍数智化共同体提供决策参考。

报告的主体包括以下三个部分：第一部分为全球数字化存量和增量调研，主要探视中国古籍的载体形态从物理的纸本到电子版、研究对象从载体到内容、数据和知识、利用方式从信息孤岛到开放互联的演进历程。尤其注重海外珍稀古籍的数字化存量和增量调研。第二部分为技术应用的范式转移评估。通过典型案例分析评估技术的迭代对中国古籍数智化开发利用范式的影响，并为应对当前 AI 技术带来的机遇和挑战提出意见和建议。尤其注重生成式人工智能、智能体、具身智能、空间智能等前沿技术如何赋能中国古籍的数智化开发利用和活化传承。第三部分为国际合作和人智协同标准的制订和应用调研分析。通过 IIIF、关联数据、MCP（模型上下文协议）等标准规范在中国古籍数智化领域的应用调研，考察国际合作的现状、分析问题和瓶颈、提出未来发展建议。尤其注重 MCP 等新型 AI 技术应用协议和人工智能伦理治理政策对中国古籍的既有开放共享模式的影响，以及可能带来的新突破、新机遇。

为了进一步明确报告的调研范围，本报告参考《古籍定级标准》（WH/T 20-2006 / GB/T 31076.1-2014）后，将“中国古籍”作如下框定：中国古籍指 1912年以前在中国书写或印刷的文献。在内容上，涵盖传统学术中经、史、子、集四部，其类型包括一般古籍、甲骨金文、简帛古籍、敦煌遗书、类书、丛书、地方志、谱牒、碑帖及古地图等。古籍是以文本、图像、实物等形式承载知识与思想，依托特定材料、制作技术与装帧形制加以定型的文化记忆媒介，在历史流传中发挥着记录、传递与累积文明成果和传承与传播文化思想的功能。中国古籍是中华优秀传统文化的重要载体，集中承载了具有中华传统的知识生产方式、文明演化历程与文化传承机制，经过有效的数智化转化后，必将在发挥重要的作用。

二 全球视野下中国古籍数字化存量与增量

随着全球文化遗产数字化进程加速，同时受到 AI 技术的影响，中国古籍在数字空间中的存在形态与利用方式正在发生一系列变化。以下是从项目类型、内容、主题领域及技术方法等方面，对遴选的 54 个案例项目与平台展开调研的情况呈现。

（一）中国古籍数字化的全球开展情况

中国古籍承载着五千年的文化记忆。自 2007 年“中华古籍保护计划”实施以来，我国古籍保护工作实现了从最初的家底不清、保护薄弱境况，逐渐向体系化与科学化发展。通过调查全国古籍普查登记基本数据库发现：截至 2024 年 10 月 27 日，累计发布 319 家单位古籍书目记录941708 条 8982920 册 [1]。由中国国家图书馆（中国国家古籍保护中心）与北京大学数字人文研究中心联合设计开发的“《国家珍贵古籍名録》知识库”平台统计显示：目前，国务院已批准公布六批《国家珍贵古籍名录》，全国已有 485 家机构或个人收藏的 13026 部古籍入选，涵盖先秦两汉至明清时期的汉文古籍、少数民族文字古籍和其他文字古籍。其中，甲骨 4 种，简帛 187 种，敦煌遗书 405 件，碑帖拓本 219 件，古地图 149 件，普通古籍12062 部（包含少数民族文字古籍及其他文字古籍）。[2]

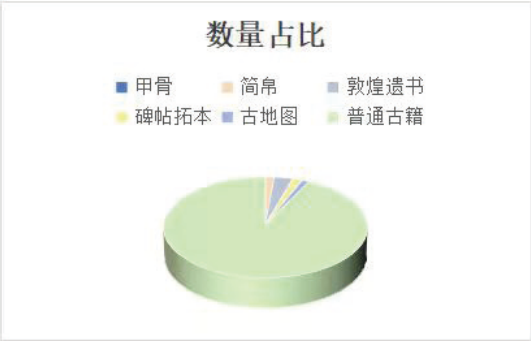

图 1 六批《国家珍贵古籍名录》古籍类型数量占比图

在普查登记工作进行的同时，数字化工作也在同步进行。截至第 11 次古籍数字资源发布，全国累计发布古籍及特藏文献影像资源将超过 16.1 万部（件）[3]。2025 年 10 月，古联公司发布的数据库产品资源规模已达 45 亿字，构成国内规模最大的线上整理本古籍资源库。[4] 由国家图书馆（国家古籍保护中心）建设的中华古籍资源库陆续发布国家图书馆藏善本和普通古籍、法国国家图书馆藏敦煌遗书、天津图书馆藏普通古籍、日本永青文库捐赠汉籍、云南省图书馆善本古籍、芷兰斋藏稿抄校本等古籍影像资源，总量超过了 3.3 万部。在古籍资源方面开放共享的举措，也迅速得到了全国各古籍收藏单位的积极响应。国家古籍保护中心联合省市区县级公共图书馆、高等院校、科研机构、博物馆系统等 40 余家古籍收藏单位在线发布古籍数字资源超过 5 万部，得到了社会各界的热烈反响和一致好评，形成了全国联动的新局面。

具体而言，目前，国家图书馆已发布馆藏古籍2 万余部，国家图书馆超过 2/3 的善本古籍实现了在线阅览。天津图书馆提供总量约 5800 余种 300 万拍明清古籍数字资源。中华古籍善本国际联合书目系统著录了三十余家海内外图书馆所藏古籍善本，数据达 2 万多条，并配有 1.4 万余幅书影。云南省图书馆提供少数民族汉文著述等古籍数字资源共 139 种 727 册给国家图书馆（国家古籍保护中心）。数字方志主要包括为清代（含清代）以前的数字方志资源 6529 种。金刻《赵城金藏》共有 1281 件数字资源已正式发布。宋人文集善本通过缩微胶卷还原数字影像并辅以详细书目，目前已发布 275 部。碑帖以国家图书馆藏有的历代甲骨、青铜器、石刻等类拓片二十三万余件为基础建设的数据库，现有元数据 2.5 万余条，影像 3.1 万余幅。甲骨实物元数据 2964 条，影像 5932 幅。甲骨拓片元数据 2975 条，影像 3177 幅。西夏文献书目数据 124 条，原件影像近 5000 拍。敦煌遗书由国际敦煌项目（International Dunhuang Programme，IDP）[5] 完成写卷影像 18 万余拍，法藏敦煌遗书完成共计 5300 余号 3.1 万余拍。谱牒部

[1] 全国古籍普查登记基本数据库 [DB/OL].(2024-10-27)[2026-03-09].http://202.96.31.78/xlsworkbench/publish.
[2] 国家珍贵古籍名录 [DB/OL]. (2024-10-27)[2026-03-09]. https://rarebib.pkudh.org/#/.
[3] 韩寒，刘彬 . 第十一次古籍数字资源联合发布——超 16.1 万部文献飞上“云端”[EB/OL]. (2026-01-06)[2026-03-09]. https://news.gmw.cn/2026-01/06/content_38520360.htm.
[4] 陈雪 . 发布古籍整理数字资源 45 亿字规模最大线上整理本古籍资源库成立十周年 [EB/OL]. (2025-10-27)[2026-03-09]. https://epaper.gmw.cn/gmrb/html/2025-10/27/nw.D110000gmrb_20251027_6-09.htm.

分，徽州善本家谱印刷资料数据库收录了中国国家图书馆藏善本古籍中徽州家谱 243 种 286 部，配有书影 5437 幅。国家图书馆与上海图书馆合作，征集该馆所藏明清家谱数字资源 2200 余种。[6]

放眼海外，中国古籍数字化仍以大型国家图书馆和高校图书馆为主体。例如，大英图书馆（British Library）通过国际敦煌项目实现了数字化敦煌文献数万页[7]。法国国家图书馆（Bibliothèque nationale de France）在 Gallica 平台上线中文古籍与手稿数千种[8]。美国国会图书馆（Library of Congress）收藏东亚文献约 4 万册，其中部分珍本已实现高清开放访问[9]。Harvard-Yenching Library 则持续推进中文古籍目录与影像数字化[10]。从整理时序来看，早期数字对象多停留在高清图像与基础元数据层面，尚未形成大规模可计算的文本资源。

不过，近年来国内数字化存量规模明显扩大、数智化加工范式正快速向人工智能转向。以国家图书馆古籍数字资源平台为代表，已累计上线古籍影像与目录数据数十万条。“荀子”古籍大语言模型项目覆盖《四库全书》等传世古籍的 40 亿字混合语料库[11]。中国哲学书电子化计划（CTEXT）收录古籍文本超过 3 万余部[12]。从本次 54 项样本统计结果看，约 70% 的项目进行了全文文本化结构化处理，约 40% 的项目进入知识图谱构建阶段，近三年新增项目中，甚至超过了一半项目与大语言模型训练直接相关。

（二）中国古籍数字化发展的阶段演进

结合调研结果来看，可以初步将中国古籍数智化进程概括为四个阶段：一是数字化阶段，以扫描、元数据规范制订、元数据著录和数字资源库建设为核心任务，此阶段解决的是数字化转换、保存与检索利用的问题。全球文化遗产数字化浪潮在新世纪前 后 进入加速期，美欧等地的数字图书馆项目、中国国家数字图书馆推广工程等陆续启动，全球文化遗产数字化浪潮推动大量馆藏古籍完成高清影像采集。二是数据化阶段，以 OCR 、自动句读、标点、分词、词性标注等技术成熟为标识。古籍文本被转化为可检索和计算的数据，多个调研案例自动句读准确率达到 90% 以上，文本识别精度显著提升，古籍数智化进入数据驱动的时代。三是语义化阶段，基本知识单元由文本向知识节点和关联发展。在此阶段，基于 BERT 模型的分类、标引、实体提取研究发展迅速。具体而言，基于知识图谱技术的知识库构建成为数智化中的主要成果，其中又以人物、地名、官职等实体抽取与知识图谱构建成为重，约 40% 的调研案例涉及知识图谱构建，语义化成为常规范式。四是智能化阶段，围绕大语言模型的垂直领域应用和模型训练成为古籍资源整合与利用的新模式。特别是 2022 年多款生成式人工智能大模型发布后，基于古籍数智化成果训练垂直领域模型成为新趋势。调研案例中多个项项目构建了垂直领域模型，语料规模达数十亿字，实现了智能化古文标引、文白翻译、智能问答与辅助研究功能。

[5] 国际敦煌项目 [EB/OL].[2026-03-09].https://idp.bl.uk/
[6] 国家图书馆 . 中华古籍资源库 [DB/OL].(2016-09-28)[2026-03-09].https://www.nlc.cn/pcab/zy/zhgj_zyk/index.shtml.
[7] International Dunhuang Programme.IDP[EB/OL].(2025-01-21)[2026-03-09].https://idp.bl.uk/.
[8] Bibliothèque nationale de France.Gallica[EB/OL].[2026-03-09].https://gallica.bnf.fr/accueil/fr/html/accueil-fr.
[9] Library of Congress.Read.gov[EB/OL].[2026-03-09].https://www.read.gov/.
[10] Harvard Library.Harvard-Yenching Library[EB/OL]. [2026-03-09]. https://library.harvard.edu/libraries/yenching.
[11] 南京农业大学。王东波教授团队荀子古籍大语言模型获中央网信办备案 [EB/OL].[2026-03-09].https://rwskc.njau.edu.cn/info/1035/4102.htm.
[12] 中国哲学书电子化计划 . CTEXT[EB/OL].[2026-03-09].https://www.wenxianxue.cn/daohang/83.html.

（三）存量与增量及趋势分析

从语料规模看,不少古籍全文资源已进入“亿字级”规模区间。尤其是大模型项目的出现，使古籍数据逐步向少数核心平台集中，呈现明显平台化趋势。具体而言，荀子古籍大语言模型项目构建了覆盖四库全书等传世古籍的混合语料库。截至 2021 年籍合网已上线资源 5000 余种，累计约 20 亿字 [13]。CTEXT 收录超过三万部文献、五十亿字符 [14]。中国基本古籍库收录自先秦至民国历代重要古籍1 万部、18 万卷、26 万卷，影像 1300 万页、录文 19 亿字 [15]。主要平台的语料库规模不断突破，甚至进入百亿字级规模。2015 年前新增项目多集中于数字化扫描与数字资源库建设，以中国基本古籍库、国际敦煌项目等为代表 [16]。2018 年后语义化智能化程度明显增强，中文古籍联合目录与循证平台的 130 万余条循证数据全部建立在本体和关联数据技术的基础之上 [17]。SikuBERT 利用四库全书繁体版本语料对 BERT 模型进行继续预训练 [18]。“吾与点”古籍智能数据与标注平台 [19] 的自动句读模型在混合文本上准确率超 94%。2022 年后智能化项目数量快速上升，本次调查样本中的荀子古籍大模型 [20]、AI 太炎 [21]、通古大模型 [22] 等多个项目构建十亿级语料规模的垂直模型。

从调研案例数据来看，中国古籍资源仍有较大的数字化发展空间。目前，在全球视野下，呈现出国内规模领先、海外珍稀突出、数智化进程不均的特点。具体而言，国内方面，上海图书馆作为馆藏规模较大、数智化工作起步较早、业界影响较大的图书馆之一，可由其古籍、家谱、碑帖三种古籍资源的数智化情况窥见一斑。其中，中文古籍联合目录暨循证平台整合 1,400 余家机构的古籍馆藏古籍书目数据，可查询 60 余万条古籍元数据记录，根据“作品 - 版本 - 单件”三层模型合并得到古籍作品 21 万余种，融合历代官修、私家、藏书楼等多种类型的古籍目录数据，形成 138 万余条循证研究数据，其中上海图书馆馆藏古籍书目 13 万 7 千余种，发布馆藏古籍扫描影像 1 万零 8 百余种 [23]。中国家谱知识服务平台可查全球华人家谱目录 9 万 8 千余种，其中上海图书馆馆藏家谱 4 万零 5 百余种，发布馆藏家谱扫描影像 1 万 4 千 6 百余种 [24]。上海图书馆馆藏碑帖约 25 万种，其中善本碑帖 3000 余种，碑帖知识库发布全文影像 45 种、单字切割图像 74198 个 [25]。在海外，哈佛燕京图书馆馆藏中文文献约 90 万册 [26]，5.3 万卷中文善本特藏已全部数字化，法国国家图书馆手稿部 [27] 的中国藏书是该部最大的馆藏，拥有超过 15 万册木刻、石印或铜版 / 铅版活字印刷的书籍，数百份手稿，以及 500 多种期刊，大英图书馆 [28] 有中文在线档案和手稿约 4000 件。

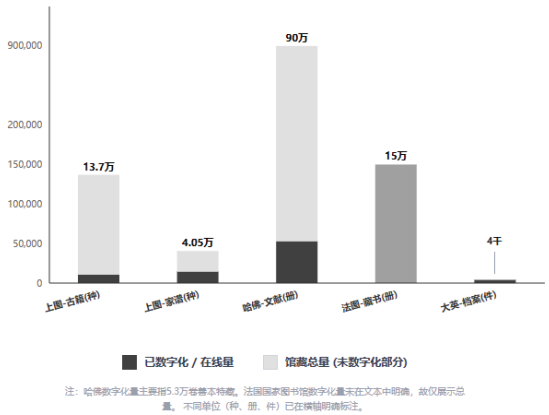

图 2 典型案例数字化情况对比图

[13] 中华书局 . 中华经典古籍库 [DB/OL]. https://www.ancientbooks.cn/helpcore?about.
[14] Chinese Text Project[DB/OL]. https://ctext.org/.
[15] 爱如生数字化技术研究中心 . 爱如生中国方志库 [DB/OL]. http://dh.ersjk.com/spring/front/news/load?id=4.
[16] 中国国家图书馆 . 中华古籍资源库 [DB/OL]. http://idp.nlc.cn/.
[17] 夏翠娟 , 林海青 , 刘炜 . 面向循证实践的中文古籍数据模型研究与设计 [J]. 中国图书馆学报 ,2017,43(06):16-34.
[18] 王东波 , 刘畅 , 朱子赫 , 等 .SikuBERT 与 SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究 [J]. 图书馆论坛 ,2022,42(06):31-43.
[19] 吾与点 [EB/OL].[2026-03-07].https://www.wuyudian.net
[20] 刘畅 , 张琪 , 王东波 , 等 . 基于大语言模型技术的古籍限定域关系抽取及应用研究 [J]. 情报学报 ,2025,44(02):200-219.
[21] 李绅 , 胡韧奋 , 王立军 . 古汉语大语言模型的构建及应用研究 [J]. 语言战略研究 ,2024,9(05):22-33.
[22] 华南理工大学深度学习与视觉计算实验室 . 通古大模型 [EB/OL].[2026-03-08].https://github.com/SCUT-DLVCLab/TongGu-LLM .
[23] 上海图书馆 . 上海图书馆中文古籍联合目录暨循证平台 [DB/OL]. https://gj.library.sh.cn/.
[24] 上海图书馆 . 中国家谱知识服务平台 [DB/OL]. https://jiapu.library.sh.cn/.
[25] 上海图书馆 . 上海图书馆碑帖知识库 [DB/OL]. https://beitie.library.sh.cn/.
[26] Harvard University Library. Harvard-Yenching Library[EB/OL]. https://library.harvard.edu/libraries/yenching.
[27] Bibliothèque nationale de France. Archives et manuscrits: Collection chinoise[DB/OL]. https://archivesetmanuscrits.bnf.fr/ark:/12148/cc3883j.
[28] The British Library. Chinese Archives and Manuscripts[DB/OL].https://searcharchives.bl.uk/?f%5Blanguage_ssim%5D%5B%5D=Chinese.

基于现有存量规模与技术升级速度，本报告对后续发展作出如下初步判断与假设：一是海外影像型数字资源与国内结构化数据对接。例如中国历史地理信息系统（CHGIS）已通过哈佛大学地理分析中心采用关联数据技术发布地名词表，提供 TGAZ API，并支持 JSON、RDF、HTML、XML 等多种数据格式返回。二是标准互通与语义协同将成为制约因素。上海图书馆、DocuSky 支持 IIIF 标准并提供 RESTful API，上海图书馆、CTEXT 提供基于 Linked Open Data 的 SPARQL 端点查询服务，CBDB 数据依据 CC BY-NC-SA 4.0 协议完全开放，但多数大模型项目暂未提供 API，识典古籍[29] 仅提供 Web 公共服务，通古大模型数据集需申请审批，开放程度参差不齐。第三，AI技术的应用在快速拉平头机构团体数智化发展差距的同时，也会扩大现存的鸿沟。

三 技术驱动的中国古籍数智化开发路径和利用模式

中国古籍数智化开发是对中国古籍文献进行数字化采集、结构化组织、语义化关联与智能化服务的系统性工程，核心目标是实现古籍的长效保存、深度挖掘与活态传承。形成了以 OCR（光学字符识别）、NER（命名实体识别）、数据库技术、知识图谱、大语言模型为核心的典型技术驱动模式。随着 AI 技术的极速发展，古籍数智化开发正处于向 AI 驱动的智能化迭代升级的阶段。

（一）古籍数智化开发的技术路线演进

经调研，本报告将古籍数智化开发总结为载体数字化、文本结构化、数据语义化、资源向量化四个层级，经历了从物理载体向数字形态转化、从可计算文本到多源知识语义关联、再到深度语义理解与智能交互的技术演进历程，逐步实现了中国古籍从可获取到可计算、可关联、可理解、可活化的层级跃升。

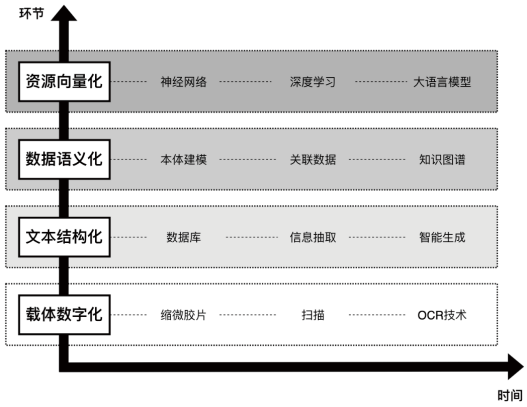

图 3 古籍数智化开发的技术路线演进

（1）载体数字化的技术路线演进

载体数字化的核心目标是将古籍从物理载体转化为数字资源，进行元数据著录，破解古籍保存难题、实现资源的可检索，扩展其可获得性。1980 年代至 2000 年代初，数字化主要依赖缩微胶片、扫描技术生成静态图像，结合人工录入与光学字符识别（OCR）

[29] 北京大学 - 字节跳动数字人文开放实验室 . 识典古籍 [EB/OL].[2026-03-08].https://www.shidianguji.com .

技术建立全文文本，同时通过扩展 GBK 字符集解决古籍用字的编码难题。这一时期，上海图书馆中国古籍善本查阅系统与中国国家图书馆的中国数字图书馆推广工程成为典型成果[30]，初步实现了古籍的可获取，但内容仍以静态图像或封闭格式为主，利用效率有限。早期 OCR 技术虽能处理部分通用印刷体识别需求，但对古籍中的异体字、生僻字识别率较低，难以满足精准数字化需求。此后，支持向量机（Supprt Vector Machine）等机器学习方法被引入，通过构造二分类器提升了字符识别的泛化能力[31]。随着智能化技术的迭代，深度学习模型成为主流 OCR 技术，卷积神经网络（Convolutional Neural Network）技术通过自动提取图像特征，大幅提升了古籍版面分析与字符识别的准确率[32]。近年来，增加了视觉层的 Transformer（ViT）模型借助自注意力机制，进一步优化了对古籍图像中的复杂版面、批注、印章的识别效果[33]。OCR 技术在 AI 的赋能下，不仅实现了异体字、避讳字的高精度识别，还具备版面重建能力，正在经历从单一文本数字化向多模态数字化的跨越。

（2）文本结构化的技术路线演进

文本结构化是将文本化古籍转化为结构化、可计算的数据形态的过程——它为后续语义化关联与智能化应用奠定基础。本世纪以来，依托数据库技术，学界与相关机构搭建了各类专题数据库，形成了多元化的实践成果。其中，哈佛大学、北京大学等合建的“中国历代人物传记资料库”（China Biographical Database，CBDB），采用关系型数据库架构，将分散于史部、子部、集部中的人物传记相关信息抽取为结构化数据[34]。马克斯 - 普朗克研究所开发的 LoGaRT（Local Gazetteers Research Tools），针对地方志文本规律性强的特点，运用正则表达式自动识别和提取重复文本模式，将文本列表转换为结构化数据表[35]。复旦大学与哈佛大学共建的“中国历史地理信息系统”（China Historical Geographic Information System，CHGIS），借助地理信息系统技术（Geographic Information System，GIS）实现了古籍中地理信息的结构化处理[36]。台湾大学项洁教授主持建设的“DocuSky 數位人文學術研究平台”，为研究者提供“一站式”文本标记与管理工具。法鼓文理學院建设的“CBETA 電子佛典集成”（Chinese Buddhist Electronic Text Association），成为佛教古籍结构化整理的典范[37]。此外，香港科技大学李中清 - 康文林研究团队与中国人民大学清史研究所共建的《缙绅录》量化数据库[38]、德龙（Donald Sturgeon）创建的“中国哲学电子书计划”（Chinese Text Project，CTEXT）[39]、北京大学朱本军主持建设的“汉语统一时间标尺平台”[40]，均是本环节的重要实践成果。在信息抽取技术方面，早期采用 SVM 等传统机器学习方法实现自动句读、命名实体识别等单任务处理，但性能有限且高度依赖人工特征标引。基于 BERT 的预训练模型（如南京农业大学、南京师范大学等联合开发的 SikuBERT）显著提升了技术性能[41]。进入 GenAI 时代后，以大语言模型为代表的技术无需复杂特征抽取，依托端到端的自然语言理解实现多任务处理，例如南京农业大学王东波教授团队开发的“荀子”古籍大语言模型，不仅支持命名实体识别、关系抽取、事件抽取等任务，还具备古籍文白翻译、自然语言问答等生成式能力，其性能提升得益于上下文学习与提示词工程的应用[42]。

[30] 赵薇 . 数字时代人文学研究的变革与超越——数字人文在中国 [J]. 探索与争鸣 ,2021,(06):191-206+232-233.
[31] 祁亨年 . 支持向量机及其应用研究综述 [J]. 计算机工程 ,2004,(10):6-9.
[32] 郭利敏 , 葛亮 , 刘悦如 . 卷积神经网络在古籍汉字识别中的应用实践 [J]. 图书馆论坛 ,2019,39(10):142-148.
[33] 顿泽栋 .“汉籍合璧”工程中古籍印章的识别与研究 [D]. 山东大学 ,2022.
[34] Harvard University.History of CBDB[EB/OL].[2026-03-07].https://cbdb.hsites.harvard.edu/history-of-cbdb.
[35] Max Planck Institute for the History of Science.LoGaRT[EB/OL].[2026-03-07].https://www.mpiwg-berlin.mpg.de/research/projects/logart-local-gazetteers-research-tools.
[36] 复旦大学历史地理研究中心 . 中国历史地理信息平台 [EB/OL].[2026-03-07].https://timespace-china.fudan.edu.cn/FDCHGIS/.
[37] CBETA 電子佛典集成 [EB/OL].[2026-03-07].https://www.cbeta.org/.
[38] 中国人民大学清史研究所 . 缙绅录 [EB/OL].[2026-03-07].http://dhiqh.ruc.edu.cn/DownloadFile/DLFile.
[39] 中国哲学书电子化计划 [EB/OL].[2026-03-07].https://ctext.org/zhs.
[40] 汉语统一时间标尺平台 [EB/OL].[2026-03-07].http://www.histchina.cn.
[41] 王东波 , 刘畅 , 朱子赫 , 等 .SikuBERT 与 SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究 [J]. 图书馆论坛 ,2022,42(06):31-43.
[42] 刘畅 , 张琪 , 王东波 , 等 . 基于大语言模型技术的古籍限定域关系抽取及应用研究 [J]. 情报学报 ,2025,44(02):200-219.

（3）数据语义化的技术路线演进

数据语义化是过去 10 年来形成的常规范式，意味着古籍数智化开发从数据整理转向知识发现，其核心特征是打破资源孤岛，实现多源异构数据的语义关联，进而支撑古籍的宏观分析与多视角可视化呈现。数据语义化的核心方法技术包括本体建模、关联数据、知识图谱等。在本体建模方面，主要分为人工（专家）构建与半自动构建。人工构建是自顶向下的，依赖领域专家定义概念、属性与关系，如上海图书馆家谱知识服务平台基于 BIBFRAME 构建家谱本体模型及词表[43]。半自动构建是自底向上的，通过从大规模数据中自动抽取出概念与概念间的关系，再辅以人工校对，形成本体模型和词表。在关联数据技术应用方面，早期的关联数据发布技术如 RDB2RDF，通过映射规则将遗留系统的关系数据库转换为 RDF 三元组，实现 SQL 到 SPARQL 的透明转换，可在不遗弃原有系统的基础上，额外增加关联数据发布层。随着文本分析、知识挖掘、实体识别、原生图数据库等技术的发展，在遗留系统之外重建关联数据发布平台成为常态。上海图书馆中国家谱知识服务平台、中文古籍联合目录暨循证平台等数字人文服务平台，通过关联数据技术，以本体建模为方法，以关联数据四原则为基本技术架构，通过将 HTTP URI 作为唯一标识符，实现不同类别古籍资源和数据的语义关联与开放数据服务[44]。与文化记忆机构注重的是关联数据技术的天然开放性不同的是，知识图谱技术更强调将分散的数据转化为可推理的语义网络，更为人文学者所重视。王兆鹏教授主持建设的唐宋文学编年地图便是典型案例[45]，该项目通过知识图谱技术与地理信息系统的深度结合，在实现唐宋文学古籍语义化的基础上，完成了知识发现与可视化呈现，为文学研究提供了全新视角。

（4）资源向量化的技术路线演进

以神经网络、深度学习与大语言模型为核心演进脉络，中国古籍开发正式进入向量化时代，实现了古籍知识的深度理解与智能交互。早期智能化实践主要依托神经网络与深度学习技术，例如清华大学孙茂松教授主持建设的九歌人工智能诗歌写作系统早期版本[46]，如上文提到的面向古汉语智能处理的预训练语言模型 SikuBERT，为古籍文本的向量化奠定了基础。2023 年以来，随着大语言模型技术的成熟，不仅主流通用大模型（如 Deepseek，Qwen 等）被逐步迁移到古籍智能处理场景，通过其强大的自然语言理解与生成能力，辅助古籍文白翻译、词义训释等基础任务，还涌现出一批基于大模型底座的古籍垂直领域大模型，如北京师范大学基于 DeepSeek-R1 开发的 AI 太炎、浙江大学徐永明教授主持建设的云四库[47]、南京农业大学王东波教授团队开发的荀子古籍大模型。这些垂直领域大模型普遍采用 LoRA（Low-Rank Adaptation）微调技术，在通用大模型基座上，依托古籍专用语料进行轻量化微调，进一步提升了古籍智能处理的效率。同时，检索增强生成（Retrieval-Augmented Generation，RAG）技术被广泛应用于古籍向量化实践，通过将古籍文本、结构化数据、语义化知识与大模型生成能力相结合，先检索数据中与提问密切相关的知识，再利用大模型生成并输出，有效提升了古籍问答、引文溯源、内容生成的准确性与严谨性，弥补了大模型易出现的知识偏差和幻觉问题。在大模型、LoRA 微调与 RAG 技术的协同赋能下，典籍资源向量化的前沿探索聚焦于智能体的实践应用，例如北京大学“吾与点”古籍智能数据与标注平台，通过设定不同角色的智能体实现古籍整理的流程化智能管理，大幅优化了古籍整理流程，推动向量化技术向全自动化、高效化方向迭代升级。得益于 GenAI 技术的发展，多模态大模型在处理古籍图像方面的能力，为新一代智能 OCR 技术的发展带来了巨大的潜力，有望实现对异体字、手写体、少数民族语言、减字谱、工尺谱等的智能识别。夏翠娟通过多模态大模型 LoRA 微调、多模态资源的向量化表示和跨模态对齐，初步实现了对古琴减字谱的识别[48]。

[43] 夏翠娟 , 刘炜 , 张磊 , 等 . 基于书目框架 (BIBFRAME) 的家谱本体设计 [J]. 图书馆论坛 ,2014,34(11):5-19.
[44] 夏翠娟 , 刘炜 , 陈涛 , 等 . 家谱关联数据服务平台的开发实践 [J]. 中国图书馆学报 ,2016,42(03):27-38.DOI:10.13530/j.cnki.jlis.160014.
[45] 唐宋文学编年地图 [EB/OL].[2026-03-07].https://cnkgraph.com/Map/PoetLife
[46] 孙茂松 . 诗歌自动写作刍议 [J]. 数字人文 ,2020,(00):32-38.
[47] 徐永明 , 王永攀 . 通用大语言模型在文史领域中的应用：以云四库智能问答系统为例 [J]. 数字人文 ,2025,(03):195-212.
[48] 夏翠娟 .GenAI 技术环境下多模态文化记忆资源的知识表示研究：以古琴减字谱为例 [J/OL]. 中国图书馆学报 ,1-30[2026-03-07].https://link.cnki.net/urlid/11.2746.G2.20251204.1826.002.

（二）典型案例的技术路径与利用模式

在典型案例分析部分，案例的选择遵循三个标准：一是覆盖面广，全面涵盖经、史、子、集四部古籍。二是技术代表性强，聚焦主流技术与创新技术的实践应用。三是实践成效明确，具备可复制、可推广的经验。表 1 所列 18 个知名项目均符合上述标准，是开展案例分析的优质样本。

表 1 典型案例调研分析

分类	项目名称	技术路径	利用模式
经	《春秋左氏传》检索平台	检索平台、人物地图	文本检索、人物时空关联浏览
史	腾讯“探元”甲骨文多模态识读项目	多模态模型、OCR、图像识别、甲骨文语义解析	甲骨文识读、考释、文物断代、文字溯源研究
	CBDB	关系数据库（Access/SQLite/SQL Server 等）、结构化提取、查询接口	传记查询、社会网络分析、GIS 整合、统计分析、历史研究
	CHGIS	GIS 平台、时间序列政区空间数据、地名查询系统	历史地理研究。多领域时空数据分析基础平台
	CTEXT	数字文献库、结构化标引、全文检索、基础统计 / 文本分析工具	全文检索。词频 / 共现等文本分析。研究与教学
	吾与点古籍智能数据与标注平台	模块化工具链（句读 / 抽取 / 标注 / 结构化）、人机协同。流程编排	整理生产：标注、校对、结构化与图谱化支撑
	中文古籍联合目录暨循证平台	MARC/DC 元数据标准化、本体知识组织、关联数据、知识图谱、题名 / 责任者识别与消歧、数智循证	目录查询、版本考证、流传分析。支撑古籍整理出版与学术研究
	中国家谱知识服务平台	本体知识组织、关联数据、GIS、RESTful API、SPARQL Endpoint	常识普及、智慧寻根、知识发现、知识挖掘
	中国地方志数字人文研究平台 LoGaRT	搜索 / 分析软件、数据库管理、搜索算法、多模态整合	历史分析、数据收集、协作研究、开放访问
	国际敦煌项目 IDP	数字化、在线数据库、多语言搜索、协作平台。人工智能增强图像分析 / 文物重建	文化保存、研究、教育资源
史	考古资料数位典藏资料库（中研院史语所）	数位典藏平台	考古资料检索与利用
子	全国古籍普查登记基本数据库网址	全国古籍元数据整合、统一检索、标准化著录	古籍普查、资源统计、馆藏查询、保护管理
	“荀子”古籍大语言模型项目	Transformer decoder-only、基于 Qwen 优化、40 亿字混合语料、古籍 - 现代汉语混合训练、智能标引 / 抽取 / 翻译 / 词法 / 标点 / 生成	古籍整理、研究、教学、诗歌生成、翻译、大众传播
	AI 太炎	参考 DeepSeek-R1 架构、1.8B 轻量、从头构建古汉语模型、训练 + 微调、释义 / 翻译 / 句读 / 用典	学术科研、基础教育、编辑出版、文化传播
	识典古籍	OCR/ 版面处理（若有）、检索增强生成（RAG）、引用回溯	研究任务辅助：资料梳理、问答、引文定位与写作辅助
	SikuBERT/ SikuRoBERTa	BERT-Base-Chinese/Chinese-RoBERTa-wwm-ext、《四库全书》繁体语料继续训练、MLM 无监督、15% 遮掩	分词、断句标点、词性标注、NER 等下游任务
集	通古大模型（TongGu-7B-Instruct）	Baichuan2-7B-Base、2.41B 古籍语料增量预训练。400 万对话指令微调、RAT 冗余度感知微调	高校 / 科研：古籍翻译与研究。句读 / 翻译 / 赏析
	中华电子佛典（CBETA）	结构化文本库、全文检索、校勘比对系统	佛教文献检索、校勘、研究、教学应用

基础资源检索类核心目标是通过便捷化技术实现古籍资源快速查阅、检索与调阅，技术特点为轻量化、便捷化、全覆盖，以检索技术、数据库建设、标准化著录为核心支撑，无需复杂智能处理。应用模式以资源整合与便捷检索为主，降低用户使用门槛。典型代表包括《春秋左氏传》检索平台，依托检索平台与人物地图技术实现经部古籍快速查阅。考古资料数位典藏资料库[49] 聚焦史部考古资料，提供便捷检索渠道。CBETA 构建结构化文本库，支撑集部佛教文献检索与基础校勘。全国古籍普查登记基本数据库通过元数据整合与标准化著录，实现四部古籍普查与馆藏查询。

[49] 中央研究院历史语言研究所 . 考古资料数位典藏资料库 [EB/OL].[2026-03-08].http://ndweb.iis.sinica.edu.tw/archaeo3/System/pages/searchSP.jsp .

中文古籍联合目录暨循证平台运用 MARC/DC 元数据标准化、本体、关联数据、数据可视化等技术，支撑古籍目录查询与版本考证，史部资源应用最突出。

学术研究辅助工具类是古籍学术研究智能化的核心支撑，通过多样化技术提供全方位服务，可细分为两类：第一类编研功能导向型工具，提供文本分析、可视化、社会网络分析等编研服务，技术特点为专业化、具象化、场景化，融合数据库、文本挖掘、GIS 等技术，应用模式为工具赋能与精准编研同时发力。典型代表有 CBDB，支撑史部人物传记编研。CHGIS 提供时空可视化服务，助力历史地理研究。CTEXT 提供文本分析工具，服务四部古籍研究与教学。第二类智能化底座支撑型工具，依托古汉语专用大模型技术提供底层支撑，技术特点为智能化、通用性、基础性，通过预训练与微调打造专用模型。典型代表有 SikuBERT/SikuRoBERTa，提供古籍基础处理支撑。通古大模型、荀子古籍大语言模型、AI 太炎等，为学术研究提供智能化底层支撑。

全流程化古籍整理类聚焦于通过流程化、标准化技术，支持海量古籍从原始文献到数字化、结构化成果的全流程处理，具有流程化、标准化、协同化的技术特征，并以模块化工具链、人机协同、元数据标准化等技术为核心。其应用模式以流程赋能与规范整理为主，实现古籍整理标准化生产。典型代表包括吾与点古籍智能数据与标注平台，提供全流程整理支撑，覆盖四部古籍。国际敦煌项目运用数字化采集、人工智能增强图像分析等技术，实现敦煌文化遗产保存与协作研究。中国地方志数字人文研究平台 LoGaRT 结合多模态整合技术，支撑史部地方志规范化整理。

古籍文化传播与知识服务类：核心是通过多元化、沉浸式技术打破传播时空限制，实现古籍文化活态传承，技术特点为多元化、沉浸式、大众化，融合多模态模型、GIS、虚拟交互等技术。应用模式以价值转化与大众传播为主，将古籍知识转化为可交互形式。典型代表有腾讯“探元”甲骨文多模态识读项目，运用多模态模型、OCR 等技术实现甲骨文识读与文化传播。中国家谱知识服务平台依托知识组织、GIS 技术，实现智慧寻根与文化普及。唐宋文学编年地图等通过可视化技术推动文学文化传播。识典古籍结合 OCR、RAG 技术，提供研究辅助服务，史部、集部内容占比集中。

上述案例应用呈现出鲜明的共性与差异化特征，形成适配不同需求的实践模式。在共性方面，其核心技术普遍采用传统数字化采集、结构化处理、检索服务等基础技术，注重资源整合与标准化建设，实现古籍从静态保存向动态利用转变，覆盖四部古籍，满足多元需求。在差异化方面，不同类别案例侧重不同：基础资源检索类侧重数字化技术，学术研究辅助类侧重深度化、智能化技术，全流程整理类侧重流程化、标准化技术，文化传播类侧重可视化、沉浸式技术。同时，经、子部侧重古汉语智能模型等技术，史部侧重 GIS、数据库等技术，集部侧重文本结构化等技术。这些特征充分反映了技术与古籍内容、应用需求的深度适配关系。

（三）古籍数智化开发利用的技术提升建议

基于上述分析，本报告针对现有实践模式的局限性与技术瓶颈，从需求适配、技术融合、标准规范、创新应用等层面提出提升路径方面的建议。

（1）数智开发利用面临的挑战

当前古籍开发在技术层面仍面临诸多瓶颈，制约开发质量与效能的进一步提升，具体主要体现在三个方面：一是技术与需求适配不足，古汉语处理的特殊性导致现有技术仍有短板，OCR 技术对模糊字迹、异体字的识别精度有待提升，大模型在中国古籍深度解读、知识准确性校验上易出现偏差，对中国古籍用典、语境的理解不够精准，RAG 技术在古籍引用溯源、上下文关联的精准度上存在不足，难以完全满足学术研究的严谨性需求。二是多技术融合深度不够，部分开发项目仍停留在单一技术应用阶段，未能实现大模型、知识图谱、GIS、人机协同等技术的深度融合，技术工具间接口不互通，缺乏全流程协同的技术体系，导致开发效率偏低、成果复用性差，难以形成技术协同与效能提升的良性循环。三是技术标准化与规范化缺失，不同机构的数智开发项目采用不同的技术标准与数据规范，如语料整理、结构化标引、元数据格式不

统一，导致不同平台的资源无法有效互通、整合，形成新的技术壁垒。四是中国古籍语料的合规使用许可缺乏和创新应用落地困难，影响大模型训练的质量与新技术应用的规范性和社会效益。

（2）数智开发利用的技术提升路径

立足中国古籍数智化开发的技术演进规律与典型实践，本报告针对当前技术瓶颈，聚焦核心技术优化与体系完善两大目标，从四方面系统提出技术提升路径，为中国古籍数智化高质量发展及古籍文化活态传承提供可行支撑。

优化核心技术与任务适配性，破解古汉语智能处理瓶颈。受古汉语特殊性影响，通用大模型在古籍处理中存在语义理解偏差、专业任务准确率不足等问题，制约开发效能。核心解决路径是推进古籍领域专用大模型研发迭代，一方面依托开源基座大模型，整合经史子集优质语料开展增量预训练与指令微调，融入人类反馈强化学习（Reinforcement Learning from Human Feedback）[50] 与多任务集成学习（Multi-task Ensemble Learning）[51]，强化模型对古汉语特殊语法、生僻字词以及文言上下文的理解，提升句读标点、词义训释、实体抽取和文白翻译等核心任务的效能。另一方面借助知识蒸馏技术（Knowledge Distill）[52] 研发轻量化专用模型，适配地方志、甲骨金文、家谱文集等不同古籍形态及移动阅读、学术编研等多元应用场景，实现“通用 + 专用”模型协同。同时进一步引入多模态特征，整合文本、书影、图表等异构数据，增强古籍数智化语义理解和感知维度，促进古籍大模型的迭代更新。

深化多技术协同融合，构建古籍活化全流程技术体系。中国古籍数智化开发利用是一项系统性工程，单一技术无法覆盖全流程需求。因此有必要推动大模型、RAG、知识图谱、GIS、人机协同等前沿技术深度融合，打通接口壁垒，构建从数字化采集到结构化处理、语义化关联和智能化服务的全流程闭环生态。具体而言，不仅利用人在回路（Human-in-the-loop）、主动学习（Active Learning）[53] 和置信学习（Confidence Learning）[54] 等技术迭代范式增强高质量数据集的产出，并反哺模型增强训练、提升基础设施的智能处理水平与综合泛化能力，而且通过嵌入多功能模块化工具链实现技术联动，设计专业人机协同机制（友好窗口、高性能功能和可视化交互）降低操作门槛，融通知识图谱、GIS 与智能问答，打造沉浸式、时空化的应用场景，提升古籍知识组织、价值显现和活化利用能力。

建立技术标准化体系，推动资源整合与开放共享。当前技术标准不统一、资源碎片化、接口不兼容问题突出，不同主体的数字化成果格式各异，导致数据无法互通、资源重复建设、利用率低下。需联合政府相关部门，高校古籍研究所、公共图书馆、古籍出版社、科技企业等多方力量，构建面向古籍数智化开发的统一规范技术标准体系，明确开发全流程的实操要求，消除技术壁垒和接口鸿沟。统筹制定中国古籍国家级、行业级标准化语料与知识基础库建设标准，规范资源管理与使用权属，明确古籍资源共享机制，为古籍数智技术的研发和优化提供高质量数据支撑，破解资源利用痛点。

强化新技术实践迭代，促进可持续性价值融创。以技术创新与实践迭代为核心，持续跟踪前沿技术并融入开发全流程，实现文化与技术价值深度融合。一方面引入多智能体协同（Multi-agent Collaboration）[55]、人智共生（Human-AI）[56] 等新型计算范式，加强多模态数据采集、处理、服务各环节的技术迭代，不断改善算法与模型性能。另一方面加快活化利用的创新实践，探索具身智能、沉浸式展演、多智能体交互等新型场景，实现技术创新与文化传承双向赋能，推动数智开发从技术落地向价值彰显的跨越，助力中华优秀传统文化在数智时代得到更好的传承，并能与 AI 技术的发展双向赋能。

[50] Chaudhari S, Aggarwal P, Murahari V, et al. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms[J]. ACM Computing Surveys. 2025, 58(2):1-37.
[51] Zhang Y, Liu T, Lanfranchi V, Yang P. Explainable tensor multi-task ensemble learning based on brain structure variation for Alzheimer’s disease dynamic prediction[J]. IEEE Journal of Translational Engineering in Health and Medicine. 2022 Nov 4;11:1-2.
[52] 黄震华 , 杨顺志 , 林威 , 等 . 知识蒸馏研究综述 [J]. 计算机学报 . 2022 Mar;45(3):624-53.
[53] Monarch RM. Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI[M]. Shelter Island, NY: Manning Publications Co, 2021.
[54] Northcutt C, Jiang L, Chuang I. Confident learning: Estimating uncertainty in dataset labels[J]. Journal of Artificial Intelligence Research. 2021, 70:1373-411.
[55] Tran K T, Dao D, Nguyen M D, et al. Multi-agent collaboration mechanisms: A survey of llms[EB/OL]. arXiv preprint arXiv:2501.06322.
[56] Amershi S, Weld D, Vorvoreanu M, et al. Guidelines for human-AI interaction[C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY: ACM, 2019: 1-13..

四 国际合作和人智协同标准的制订和应用

在 AI 技术引发的全球科研范式变革中，中华典籍的保护与传承正进入一个以全球协作治理与底层标准深度融合为特征的新周期。由于历史原因，海量中国古籍广泛散佚于全球各大文化记忆机构。这种物理分布的碎片化、标准规范应用的不均衡，直接导致了数字空间中数据孤岛的长期并存。因此，制订并推广统一的互操作标准，不仅是实现跨国界资源调阅的技术前提，更是人工智能时代构建权威、可信、可解释的数据底座的制度保障。通过对古籍数智化开发利用相关的标准规范应用调研，深度解析国际图像互操作框架（International Image Interoperability Framework，IIIF）[57]、关联数据(Linked Data)[58]等成熟国际标准的现阶段应用成效。同时，前瞻性地探讨模型上下文协议（Model Context Protocol，MCP）等新兴 AI 协议与伦理治理政策在古籍活化中的战略意义，旨在为构建“人智协同”的全球中华典籍数智共同体提供决策参考。

（一）数字化基座的全球协同：国际通用标准应用调研

在迈向数智化深度开发利用之前，全球范围内对中国古籍的数字化治理已形成以底层协议为核心的协同网络。这种协同不仅是存储介质的转换，更是通过 IIIF 与关联数据（Linked Data）等国际标准，重构了古籍在数字空间的交互逻辑与语义关联。

（1）国际图像互操作框架（IIIF）：跨机构资源的虚拟整合与模式演进

调研显示，国际图像互操作框架（IIIF）已成为全球头部记忆机构处理中国古籍图像资源的事实标准。其核心贡献在于通过标准化的 Image API 与 Presentation API，解决了跨馆藏资源在同一逻辑层面的调阅与比对难题。从调研案例来看，目前已形成了两种代表性的应用模式：

一是“集中式基础设施模式”，以哈佛大学图书馆为代表。哈佛图书馆技术服务中心（LTS）构建了统一的 IIIF 传输底座，支持哈佛燕京图书馆将 13 世纪至 19 世纪的中华善本、地方志等数千部珍稀资源全面接入。通过采用 Mirador 作为主要查看器，允许研究者在一个窗口内调阅哈佛馆藏，支持多窗口比较、深度缩放及图像标注，为美国高校东亚图书馆数字化项目树立了标杆。[59]

二是“跨国多边协议模式”，以国际敦煌项目（IDP）为典型。该项目由英国国家图书馆发起，涉及中国国家图书馆、敦煌研究院、日本龙谷大学、俄罗斯科学院东方文献研究所、法国国家图书馆等全球 35 家合作机构。利用 IIIF 标准，IDP 将分散在世界各地的敦煌文献（写本、绘画、纺织品等）在单一在线平台上进行数字化重组，成为开放获取和国际合作的典范项目。[60]

国内机构的实践如上海图书馆的碑帖知识库与复旦大学的印谱文献虚拟图书馆，则展现了 IIIF 在垂直细分领域的深度应用[61]。上海图书馆通过发布 IIIF 图像 API，支持深度缩放与区域裁剪，显著提升了金石拓片等微观纹理的研究支持效果。调研也发现，IIIF 的推广仍面临边际成本递增的困境：一方面，建立和维护图像服务器、存储系统需要持续投入资金，技术维护成本较高。另一方面，对于非书册型古籍（如长卷、简牍、手稿等特殊格式），现有的呈现逻辑往往难以精准还原其物理叙事逻辑。

（2）关联数据（Linked Data）：知识实体的结构化重塑与开放语义互联

如果说 IIIF 实现了古籍数字资源对象层面的互通，关联数据（Linked Data）则从语义层面实现了古籍知识的解构与重组。利用本体、RDF（资源描述框架）、与 HTTP URI 唯一标识符，古籍内容从封闭的文本块演变为可计算、可链接、可开放获取的知识节点。本报告在调研时重点关注了几个具有典范意义的项目。

在人物关系网络维度，CBDB 通过对史部、子部与集部中与人物传记关联的部分进行结构化处理，为数

[57] https://iiif.io/
[58] https://www.w3.org/DesignIssues/
[59] Harvard Library. Chinese Collection[EB/OL].[2026-03-08]. https://library.harvard.edu/collections/chinese-collection.
[60] British Library. The International Dunhuang Programme[EB/OL].[2026-03-08].https://idp.bl.uk.
[61] 程静 . 基于国际图像互操作框架的数字特藏资源建设研究 [J]. 数字图书馆论坛 ,2022,(04):11-16.

十万历史人物赋予了唯一的语义标识。这种关联数据的发布，使得研究者能够利用 SPARQL 查询，在秒级时间内完成跨地域、跨时代的社会网络演化分析。[62]

在时空维度，CHGIS 历经二十余年建设，已构建起覆盖 2000 年中国历史行政区划的时空框架与统一数据标准，将《二十四史》与历代地方志中的地名、行政区划进行语义化 [63]。进入 AI 时代，复旦大学历史空间综合分析实验室依托开源大语言模型，以智能替代考据为目标，实现从方志、政书等海量古籍中自动识别、抽取并时空锚定历史地名。三方面协同发力，推动历史地理学从文献整理和 GIS 呈现迈向 AI 驱动的知识生成、动态推演与系统模拟 [64]。

在全文语义化层面，CTEXT 全面覆盖经、史、子、集，展示了关联数据与 OCR 技术的深度结合 [65]。上海图书馆的实践进一步拓宽了边界，中国家谱知识服务平台等数字人文平台采用基于 BIBFRAME 的统一知识建模和基于 RDF 的一致知识表示，使图书馆从传统的文献提供者转变为以数据和知识为基本信息单元的服务提供者。

但与此同时，全球协同中本体异构的问题依然突出：相对于海量中国古籍，已实现关联数据化的仍是极小部分。由于缺乏统一、被广泛接受的中国古籍顶层本体模型，不同项目在描述同一历史逻辑时往往采用不同的概念定义，导致语义关联在跨库检索时仍存在严重的逻辑断裂。

表 2：中国古籍国际标准规范应用情况调研

机构名称	项目	古籍类型	跨馆协同	关联数据	IIIF
上海图书馆	中国家谱知识服务平台	谱牒		✓	
	中文古籍联合目录暨循证平台	官修、私家、史志、藏书楼、版本目录书	与柏克莱加州大学东亚图书馆、哈佛燕京图书馆、澳门大学、美国哥伦比亚大学东亚馆等合作	✓	
	碑帖知识库	碑帖			✓
北京大学、哈佛大学等	中国历代人物传记资料库（CBDB）	史部、子部等与人物传记关联的部分	由北大、哈佛等合作建设，其关联数据由上海图书馆开发维护	✓	
复旦大学、哈佛大学	中国历史地理信息系统（CHGIS）	史部为主	由复旦大学、哈佛大学于 2001 年启动建设	✓	
哈佛大学哈佛燕京图书馆	中文善本特藏数字化（Harvard-Yenching Library digitized version of Chinese Rare Books）	13 世纪至 19 世纪善本，明清时期个人著作，家谱、地方志等	2009 年国家图书馆与哈佛大学哈佛燕京图书达成协议，对哈佛燕京图书馆所藏中文善本和齐如山专藏进行数字化		✓
跨机构合作	国际敦煌项目（IDP）	分散在全球的各类敦煌文献	英国国家图书馆、中国国家图书馆、敦煌研究院、日本龙谷大学、俄罗斯科学院东方文献研究所、法国国家图书馆等 35 家合作机构。		✓

[62] Harvard University.CBDB Linked Open Data[EB/OL].[2026-03-08]. https://chinesecbdb.hsites.harvard.edu/cbdb-linked-open-data
[63] 复旦大学历史地理研究中心 . 中国历史地理信息平台 [EB/OL].[2026-03-08].https://timespace-china.fudan.edu.cn/FDCHGIS/.
[64] 复旦大学历史空间综合分析实验室简介 [J]. 历史地理研究 , 2020, 40 (02): 162.
[65] 中国哲学书电子化计划 [EB/OL].[2026-03-08]. https://ctext.org/tools/zh.

（二）人工智能环境下的标准制订：前沿趋势与治理逻辑

随着大语言模型（LLM）成为数字人文研究与实践的新型基础设施，传统的静态标准已不足以应对GenAI 带来的挑战。国际学术界正将注意力转向如何构建可信、可溯源的智能化古籍数智化体系。

（1）MCP 协议：构建智能体时代的知识主权与溯源机制

在大模型应用中，最为严峻的学术伦理问题是事实幻觉。由于大模型本质上是基于概率分布的预测，缺乏对确定性事实的逻辑计算，在解读中国古籍时极易产生伪造引文等幻觉。在传统的资源集成模式中，由于数据孤岛的长期存在，开发者必须针对不同数据源编写特定的 API ，容易造成 N 个 AI 模型和 M 个异构数据库的组合爆炸问题。

为此，由 Anthropic 等前沿 AI 机构提出的模型上下文协议（MCP），为古籍数智化提供了新的路径。MCP 采用客户端 - 主机 - 服务器的架构，其核心在于规范了不同数据与不同模型之间的统一的接口，定义了智能体（AI Agent）与权威数据库之间的动态交互标准 [66]。

通过 MCP 协议，大模型在生成关于古籍的叙事时，不再仅仅依赖概率预测，而是通过标准化接口实时调阅经过点校的全文库、高精度的图像资源库和高质量的语义知识库。这种模式实现了模型生成与事实检索和语义逻辑的深度解耦，确保了 AI 引文的精确溯源，有望解决了大模型在处理晦涩古籍文本时因推理过度而产生的幻觉问题。

同时，MCP 协议亦为数字主权保护提供了技术手段：古籍服务机构可以通过协议规定 AI 对数据的使用粒度，在不泄露核心敏感数据的前提下，实现古籍知识的受控输出。这标志着国际合作从简单的数据开放转向了更高阶的智能对接。

（2）AI 伦理治理：算法公正性与文化敏感性的双重博弈

AI 在古籍处理中的广泛应用，引出了深层的伦理风险。运用 Mepham 的伦理矩阵从利益相关者角度进行分析 [67]，当前面临以下三大危机：

首先是知识幻觉与历史虚无主义的潜在威胁。若训练语料中蕴含特定时期的意识形态局限或刻板印象，AI 生成的历史解读可能加剧文化偏见。模型甚至可能杜撰从未存在的史料，消解历史的严肃性。

其次是算法黑箱与解释权的去权威化。普通用户无法判断 AI 给出的古文解释是否基于权威注疏，算法可能在无意识中摒弃了学术传统，造成特定历史偏见的二次扩张。正如 Safiya Noble 在《压迫算法》（Algorithms of Oppression）中所警示，技术并非价值中立。

最后是数据污染与数据掠夺的困境。互联网上充斥的低质量、未经审校的 AI 生成内容，正逐步稀释权威古籍数据的权重。同时，商业大模型通过网络爬虫等技术手段，无偿抓取公共文化机构的数智化成果进行模型训练。这种数据掠夺侵害了整理者的劳动权益，切断了数据生产者与收益分配之间的关联。

针对上述风险，构建针对文化遗产的 AI 伦理治理框架势在必行。在数据溯源层面，参考 CIDOC CRM 模型，建立分级数据呈现机制，每一条知识关联都应具备可追溯性 [68]。在算法逻辑层面，必须将人重新置于关键节点，引入强人工反馈（RLHF）机制，让典籍专家参与算法的价值对齐过程，在知识库中保留争议性的学术观点，防止算法的专断。

同时，针对跨境合作中的数据安全问题，联邦学习技术展现出巨大潜力——即实现在数据不出境、不泄露隐私的前提下，多国机构共同训练具有全球视野的中国古籍大模型。该模式将有效平衡资源开放共享精神与国家文化数据安全。在政策倡导上，需结合国家《关于加强科技伦理治理的意见》[69] 及《生成式人

[66] 阿里云开发者社区。一文掌握 MCP 上下文协议：从理论到基础 [EB/OL]. (2025-04-03)[2026-03-07].https://developer.aliyun.com/article/1659874
[67] Mepham B. A framework for the ethical analysis of novel foods: The ethical matrix[J]. Journal of agricultural and environmental ethics, 2000, 12(2): 165-176.
[68] https://cidoc-crm.org/
[69] 中共中央办公厅，国务院办公厅。关于加强科技伦理治理的意见 [EB/OL]. (2022-03-20)[2026-03-07]. https://www.gov.cn/gongbao/content/2022/content_5683838.htm.
[70] 国家互联网信息办公室，国家发展和改革委员会，教育部，等。生成式人工智能服务管理暂行办法 [EB/OL]. (2023-07-10)[2026-03-07]. https://www.gov.cn/gongbao/2023/issue_10666/202308/content_6900864.html.

工智能服务管理暂行办法》[70]，遵循人机协作共生理论、数据隐私最小化原则与文化原真性伦理。

（三）行动倡议：构建人智协同的全球性中国古籍数智化共同体

中国古籍的数智化演进已跨越技术积累阶段，进入到标准引领与治理制衡的新周期。为应对这一变革，本报告提出以下行动倡议：

建立中国古籍通用本体。联合全球文化记忆机构与 DH 研究中心，制订一套跨越语言障碍、兼容古今地理与职官体系、涵盖人物与历史事件、支持各种历史纪年映射的的本体模型和语义编码标准，为全球范围内的中国古籍融合扫除障碍。

探索基于 MCP 协议的 AI 基础设施建设。共同开发面向大模型的标准化数据调用接口规范，确保全球研究者在使用 AI 处理中国古籍时，能够接入可信任、可验证的数据源。

制订《中国古籍数智化伦理公约》。建立一套涵盖算法透明度、历史真实性保护及数字主权保障的伦理准则，防止 AI 技术在活化利用过程中的文化消解与数据误用。

在全球视野下，可通过底层协议的标准化、语义层面的深度耦合以及治理逻辑的科学制订，以确保中华古籍在智能时代不仅是静态的历史记忆，更是能够被全球文明共同计算、理解与传承的智慧资产。

表 3：人智协同下中国古籍全球治理与技术应用演进蓝图

维度核心	技术标准与协议	伦理治理与约束机制	协作模式演进
资源与数据基座层	IIIF (图像互操作框架)。LOD (关联开放数据)。TEI、DC 元数据	保障数据完整性与安全性。明确版权与收益归属，抵御数据掠夺	跨国多边协议模式。集中式基础设施模式
算法与交互层	MCP (模型上下文协议)。联邦学习技术	破解算法黑箱与解释权去权威化。解决大模型“事实幻觉”与数据污染	全球智能体互联。数据不出境前提下的多国共训
人文与制度层	CIDOC CRM 数据溯源模型。专家纠偏与引用标准重构	人机协同的强化学习 (RLHF)。遵循文化原真性与数据隐私最小化原则	制定通用本体与《全球伦理公约》。专家众包校勘审核

五 结语

中国古籍作为中华民族五千年文明薪火相传的载体，不仅是历史的记录媒介，更是文化基因的传承载体。在 AI 技术极速发展并深度重塑社会生产力的今天，中国古籍的保护与传承，已不再局限于传统的数字化扫描、长期保存和检索查阅，而是迈向了以资源、数据和知识为基石、以智能技术为引擎的数智化新纪元。本报告通过对全球范围内中国古籍数字化存量与增量的调研，试图勾勒出中国古籍从静态的纸本和封闭的孤岛走向开放互联和智能利用的数字生态演进轨迹，尤其是海外珍稀古籍的开放共享和跨网域链接，彰显了数智技术跨越时空阻隔的强大力量和中华优秀传统文化的凝聚力。

技术的范式转移为本报告的核心理念提供了坚实支撑。生成式人工智能、智能体等前沿科技的涌现，不仅极大地提升了古籍整理的效率，更从根本上改变了与古籍对话的方式。古籍不再是静止的文本，而是可交互、可推理、可再生的智慧数据。通过深度挖掘经史子集背后的知识脉络和文化基因，数智技术让古老的智慧在现代语境下焕发新生，实现了从故纸堆到活知识的质的飞跃，降低了利用的门槛。然而，机遇与挑战并存，技术的迭代也要求我们在伦理治理、版权保护及算法偏见等方面保持高度的清醒与审慎，确保技术始终服务于文化的本真传承。

构建全球性的中国古籍数智化共同体，是本报告撰写的最终愿景。在 IIIF、关联数据以及模型上下文协议（MCP）等国际标准的推动下，跨国界、跨机构的合作已成为可能。标准的统一打破了数据壁垒，有望促进人智协同的深度发展，为全球学者提供了一个共建、共享、共创的广阔平台。未来，我们需要进一步深化国际合作，制定兼具前瞻性与包容性的技术标准与伦理规范，让中国古籍在数智社会的洪流中，既能保持其独特的文化主体性，又能以开放的姿态融入全球文明交流的浪潮。

中国古籍的数智化转型是一场关乎文明赓续的深刻变革。它要求我们以科技为翼，以文化为魂，在守护传统与创新未来之间找到最佳平衡点。尽管技术飞跃显著，但全球数智化进程仍面临算法偏见、算力分配不均及学术伦理等挑战。一个前所未有的人机关系重构的奇点正在临近，技术正在多方位赋能古籍数智化的全流程，而古籍数智化的成果也必将为 AI 技术的可持续发展保驾护航。为此，中国人民大学数字人文研究院发出倡议，开展全球中国古籍数智共建、共享、共创，推动中国古籍数字资源库、语料库和知识库建设融入 AI 技术的发展议程，为技术的发展构筑知识和伦理的护城河。同时呼吁全球学界、文化记忆机构与技术界跨界协作，让深藏在文化记忆机构中的典籍在算法与智慧的加持下，成为碳基人类和硅基智能共同的文化记忆，助力 AI 技术的良性发展。