高校图书馆手稿数字化与在线揭示的实践研究——以华东师范大学图书馆手稿数据库建设为例*

张　毅

（华东师范大学图书馆　上海　200241）

［摘　要］国内高校图书馆收藏了大量的纸质手稿，但由于缺少数字化、长期保存以及在线揭示的整体解决方案，无法有效发挥其价值。本研究在手稿数据库需求分析的基础上，以华东师范大学手稿数据库建设为例，利用开源软件 Goobi 实现分布式手稿加工与发布，借助人工智能技术对手稿进行深度挖掘，基于 IIIF 的底层架构使特藏库可以整合国际知名文化遗产机构发布的手稿资源到本地。根据手稿资源的特点，构建手稿在线阅读图像工具、注释转录、手写识别、特征识别、自动标签等功能，提升用户体验。未来还将不断探索名人角色扮演参考咨询、多光谱三维手稿浏览以及手稿鉴定等功能，推动国内高校图书馆手稿数据库的建设。

［关键词］高校图书馆　手稿数据库　Goobi　分布式工作流　ChatGPT

［中图法分类号］G250.74

［引用本文格式］张毅 . 高校图书馆手稿数字化与在线揭示的实践研究——以华东师范大学图书馆手稿数据库建设为例 [J]. 图书馆，2024（5）:84-92.

基于订阅购买的数字资源服务模式，使得任何高校图书馆只要有充足的经费，就能在短期内建设海量的数字馆藏，这种模式存在严重的资源同质化困境[1]，对图书馆收藏的稀有且高价值的手稿进行数字化与在线揭示，能够有效化解这一矛盾[2]。比如剑桥大学数字图书馆将手稿作为其重要收藏，其中达尔文手稿、达尔文与胡可的通信、牛顿手稿、希伯来语手稿等各类手稿占比很高，甚至还收藏了明清时期的官方公告、科举试卷等中国古代手稿[3]。哈佛大学数字图书馆收藏有欧洲中世纪手稿、莫扎特音乐手稿以及中国纳西族手稿等[4]，芝加哥大学数字图书馆收藏有肖邦的音乐手稿、奴隶制契约、林肯信件等资源[5]，吸引了大量学者前去研究，其中耶鲁大学数字图书馆发布的伏尼契手稿更是引起全球研究人员的广泛关注[6]。

国内高校图书馆也非常重视手稿特藏资源的建设，本研究于 2023 年 4 月通过图书馆网站对国内 42 所“双一流”高校图书馆的手稿收藏进行了调查，除了 1 所图书馆官网主页无法打开，其中 29 所图书馆有纸质手稿收藏的介绍，占比约为 70%，比如复旦大学图书馆的“卿云书房”与上海交通大学的“李政道数字资源中心”等[7-8]，华东师范大学图书馆于 2022 年 11 月启动了虚拟手稿馆的建设。虽然国内高校图书馆在手稿资源收藏方面已经做出了非常大的努力，然而并未像国际知名高校图书馆那样，大规模推动手稿的数字化、数据化建设，导致这些具有重要历史、学术、文化价值的手稿，只能躺在图书馆的库房中，无法充分发挥其价值。在调查的 42 所高校中，仅复旦大学、上海交通大学以及华东师范大学的图书馆发布了关于馆藏手稿的展览网站，北京大学图书馆虽然已经数字化了部分手稿，但并未对外发布[9]。手稿展览网站仅在一定程度上宣传了馆藏的手稿，但还不能起到支持教学科研的作用。

为了能够充分发挥手稿特藏资源的价值，本研究将借

* 本文系国家社会科学基金项目“高校图书馆特藏资源服务模式及站群系统研究”（项目编号：21BTQ100）研究成果之一

助开源软件与云开放平台，构建适合高校图书馆手稿资源的全流程解决方案，是一套包括数字化、元数据著录、高清发布、集成共享以及在线研究的开源解决方案。

1 手稿概述

1.1 手稿的定义

总结剑桥大学图书馆达尔文手稿（Darwin Manuscripts）[10]、哈佛大学图书馆的纳西族手稿（Naxi Manuscripts）、巴伐利亚图书馆手稿收藏（Manuscript Collections）[11]，手稿可定义为作者亲笔书写的稿本、签名、书信、书法、日记、绘画、会议记录以及抄本等，通常可分为文学手稿、绘画手稿、音乐手稿、翻译手稿。“manuscript”作形容词时意为“手写的”，作名词时又有“稿件”之意，本研究的手稿是指采用非印刷技术的亲笔所写材料，更加准确的手稿翻译，应该是“hand written manuscript”，也包括亲笔签名（autograph）与亲笔文件（holograph）。

1.2 手稿的重要价值

手稿的笔迹包含着作者的情绪、书写风格以及作品的形成过程，能让读者走进作者的精神世界[12]。手稿中包含着具体而生动的历史细节，能够还原历史现场，从手稿的笔迹风格可以了解到书写方式的时代风貌与流变，从手稿的用纸、材料、造型等，可以看到时代风尚、物质生活面貌与生产方式。

1.3 手稿的特点

一是脆弱。相较于印刷资料，手稿所使用的笔与纸张并非专业材料，比如圆珠笔、铅笔、草稿纸、报纸等材质，它们很容易在短时间内损坏消失。

二是分散。高校图书馆的手稿一般是通过校友捐赠的方式获得，不仅在空间上分散，时间上也非常分散，图书馆常常需要长期接受捐赠，才能形成一定规模的收藏。

三是稀缺。数字技术引发的书写方式和出版环境变革，使得内容创作海量涌现，但是手稿总量却在不断减少，高价值手稿更是稀缺。

1.4 研究现状

笔者于 2023 年 5 月在知网的期刊数据库中以关键词“手稿”进行检索，筛选出 2 864 篇相关的 CSSCI 期刊论文，发文量呈逐年上升趋势，涉及的主要研究领域有马克思主义、哲学、文学、图书馆学、历史学、美术等，南京大学、北京大学以及中国人民大学的发文量均超过 100 篇，国家社科基金支持了 18% 的研究成果。

虽然有关手稿的研究在不断增加，却鲜有关于手稿数据库建设的研究。当前的研究大多还停留在构思、规划阶段，比如，陈思航从手稿的元数据、文本化、规范控制等角度分析了手稿库的建设规划，但并未给出具体建设实践[13]；刘荣弟等以“饶余燕音乐手稿数据库”为例，构思手稿库的建设方案，同样未有详细的实践过程[14]；陈以敏等从手稿的收集整理、数据加工、标准规范、版权保护等方面对手稿数据库的建设提出设想，提出建设手稿数据库存在缺乏资金、技术、标准等难题[15]。

2 手稿数据库软件系统需求分析

手稿数据库由图书馆组织建设运维，面向全校师生提供服务，具有支持教学科研，传承本校文脉，以及保护国家文化遗产的作用[16]。图书馆、师生以及所在高校对手稿数据库都有着不同的需求，下面分别进行阐述：

2.1 满足图书馆资源建设需求

开源、免费、具备强大社区的数字资产管理平台可以满足图书馆对手稿数据库建设的基本需求，除此之外，还需满足以下要求：

2.1.1 能够实现数字手稿长期保存

作为手稿数据库的第一负责人，确保手稿资源的安全与长期保存，是图书馆的核心任务。除了在硬件层面采用专业存储、异地备份、网络防火墙等措施外，在数据层对出错的数据进行校验与恢复也必不可少，同时还需关注硬件、软件、数据格式过时可能导致的数据损坏[17]。

2.1.2 数字化与元数据加工

手稿资源需要图书馆长期积累，不断数字化而形成，所以图书馆需要能够方便高效对手稿数字对象与元数据进行管理的工具，这个工具还需具备多人协同处理、专业元数据词表、支持关联数据以及导入导出等功能，最好可以将本地加工好的数字对象与元数据一键发布到手稿数据库，并且与手稿数据库中的词表、集合、站点一一对应。

2.1.3 在线发布

在线发布手稿是手稿数据库的基本功能，发布平台需要具备强大的词表管理能力、数字对象动态发布能力、手稿数据管理能力、数据导入导出功能，无需任何编程技术，以可视化的方法设计数据库的发布页面、分类浏览、高级

检索等。

2.1.4 互操作能力

互操作是指数据库既能共享本地资源，又能嵌入外部资源[18]，在数据层使本地手稿库成为全球特藏资源的一部分，借鉴哈佛大学、剑桥大学、芝加哥大学等高校的数字图书馆所采用的大规模数字对象在线发布与互操作解决方案，使本地手稿资源融入其中，既能丰富本地手稿库资源，又能推广本地手稿库。

2.1.5 自动生成元数据

已经较为成熟的手写识别技术，可以帮助图书馆对近代手稿进行全文识别，再借助生成式人工智能技术，从识别的全文中自动生成 Json 格式的元数据[19]，从而更好地对手稿进行分类，识别的全文还可以实现全文检索、作者情绪分析等功能。除此之外，还需要利用人工智能技术自动识别手稿中的印章、签名等特殊内容，以便于更好地挖掘手稿价值。

2.2 满足师生对海量手稿数据发现的需求

2.2.1 拥有海量稀缺且高价值资源

本校馆藏的手稿资源，相比于互联网上浩如烟海的手稿资源，只占极小部分。所以手稿数据库需要具备集成其他平台手稿的能力，比如可以将剑桥大学的牛顿手稿、芝加哥大学的林肯手稿以及巴伐利亚图书馆的西方手稿等，经过分类整理之后，整合到本地系统。为师生提供上传手稿的功能，通过一定的激励策略，扩充手稿数据库的数据量。与此同时，还可以将手稿数据库打造成为一个站群平台，任何需要手稿数据库软件系统的本校师生，向图书馆申请后就能获得一个完整手稿数据库的网站，从而构建所在高校的手稿数据库平台，平台可对所有站点的手稿进行集成揭示，从而使手稿数据库的资源量不断扩大。

2.2.2 资源可被发现

读者是否可以轻松地发现所需资源，是手稿数据库价值发挥的重要基础。这就要求手稿数据库除了具备分类浏览、标签聚类浏览、时空可视化浏览、全文检索、知识图谱等功能，还要根据用户特点，动态推荐内容给师生，比如可以根据访问手稿库读者的专业、浏览记录、图书借阅记录等进行推荐，使手稿与读者需求深度匹配。

2.2.3 提供支持在线教学与科研的工具

在线标注与转录是对手稿资源进行研究必不可少的工具，还需要强大的用户管理能力，以实现对标注与转录的数据的保存或导出，同时还可以收藏用户关注的手稿。为了更加仔细地研究手稿的细节，还需提供图像深度缩放、对比度、亮度、锐化、旋转、色相更改等图像增强功能[20]。如果条件允许，图书馆最好提供手稿的 3D 模型，并提供在线紫外线手电筒工具，可以发现某些手稿中包含的特殊细节。

2.2.4 智能咨询与推荐

通过将手稿库中的数据导入类似于 ChatGPT 这样的大型语言模型中，实现手稿数据库内容的呈现。读者可以通过自然语言方式与手稿库进行交互，向数据库提出问题，查询文献、翻译、分析和解读手稿内容。在人工智能技术的帮助下，手稿库能够自动识别相关的文本信息，智能推荐与用户查询相关的资料和文献，为师生提供更加个性化的服务。同时，基于人工智能技术的自动标注和分类能力，可以大大提高手稿数据库的利用价值和效率。

2.3 满足高校文脉传承与吸引人才的需要

2.3.1 成为高校与名师的精神纽带

手稿数据库中保存的名师日记、笔记、专著、书信等手稿，承载着高校与名师之间的精神纽带[21]，记录了名师在高校工作学习的历程、学术成果以及生活态度，是名师与高校之间不可分割的关联。当名师自己、家人、学生或朋友在手稿库中看到这些生动的手稿时，一定能够深刻地感受到名师创作时的心境、思想和情感，更进一步增强对其所在高校的认同感。比如华东师范大学数学系张奠宙教授与杨振宁先生关于数学问题的书信手稿，见证了他们之间的深厚友谊，也为华东师范大学率先出版《杨振宁文集》奠定了基础[22]。

2.3.2 名师的精神不断激励全校师生

名师手稿是高校文化传承的重要组成部分，它们为高校师生提供了一扇窥视名师生活、思想和文化的窗口，让师生从中受益、学习，并将这种精神遗产代代相传。比如华东师范大学吕思勉先生的《文心雕龙札记》手稿中，有一部分并非吕思勉先生亲手所写，而是其父亲代写，这是由于吕思勉在写作的高潮时期突发眼疾所致，尽管如此，他依旧尽最大努力完成学术著作的撰写。胡焕庸先生的部分手稿笔迹突然大变，写得歪歪扭扭，这是因为晚年的胡焕庸先生右手颤抖严重，为了继续坚持创作，为国家的发展贡献力量，他改用左手创作。还有童世骏先生厚厚的《马克思主义哲学》讲稿、刘永翔先生《清波杂志校注》底稿、曹锡华先生《复半单李代数》译稿等皇皇巨著手稿，从他们反复修改的批注中，能够感受到这些学者对作品千锤百炼和殚精竭虑的精神。

2.3.3　吸引人才

大量稀缺且高价值的第一手手稿资源，是研究人员开展某些研究的先决条件，研究人员必须到收藏手稿的高校，才能获取这些稀缺的资源。比如斯坦福大学胡佛研究所图书馆和档案馆保存的蒋介石手稿[23]，芝加哥大学的林肯信件手稿，成为研究某些历史人物的基础，上海交通大学地方历史文献数据库保存的大量民间契约、账簿、诉讼文书等手稿[24]，是研究 18 世纪以来中国乡村经济、法律的第一手资料，上海图书馆收藏的国内关于莎士比亚研究的手稿，是学者研究莎士比亚中国形象发展的重要资料[25]。王元化先生将自己的手稿以及其他研究材料捐赠给学校[26]，后代学人想要继续完成王元化先生生前未完成的重要课题，就必须来到华东师范大学。

3　手稿数据库建设实践

本研究以华东师范大学手稿数据库的建设为例进行实践研究，探索建设一套基于工作流的分布式手稿数字化加工、元数据库著录、长期保存、在线发布以及支持教学科研的完整手稿数据库建设方案。华东师范大学图书馆非常重视特藏资源的建设，尤其是手稿特藏，经过长期的采访收集整理，已经形成了较为体系化的手稿资源，并于 2022 年 11 月成立手稿馆，同时举办“积健为雄——华东师范大学学人手稿展”活动[27]，极大地推动了手稿资源的建设。图书馆在收集手稿资源的同时，也在不断地进行手稿的数字化加工，希望以数字化的方式发布手稿资源，支持学校的科研与教学。

3.1　使用 Goobi 数字化工作流管理与发布系统

图书馆的手稿资源是多位馆员长期收集整理，并不断进行数字化的成果，数字化过程会产生大量的图片，还需要管理这些图片对应的元数据，为了适应这个数字化建设模式，更好地管理这些数字化后的图片与元数据，跨越数据库建设过程中的技术门槛，本研究采用德国 Intranda 公司开源的 Goobi（goobi.io）系统构建手稿数据库，它是一款基于 Java 开发的 Web 分布式数字化工作流管理、数字对象保存与发布的工具集成，Goobi 主要由 Goobi-workflow 与 Goobi-viewer 组成。本研究使用 2 台 Ubuntu 20.04 的虚拟机分别部署这两个应用，使用 JDK、Tomcat、MariaDB、Apache 以及 Apache Solr 等开源工具来运行它们。

3.2　手稿数字资源在线发布

3.2.1　基于工作流的馆藏纸质手稿数字化

Goobi-workflow 主要面向数字资源加工人员，它以工作流的方式将众多地理位置分散的手稿数字化人员组织在一起，共同完成纸质资源的数字化，同时以可视化的方式了解每个数字化人员的工作进度，经过数字化的手稿可以自动同步到 Goobi-viewer 中，也可以导出为需要的格式进行长期保存，或者发布为 REST API 对外服务。华东师范大学图书馆利用 Goobi-workflow 将手稿数字化分解为准备扫描的纸质手稿、初始数据录入、扫描人员进行数字化、OCR 人员进行文字识别、手稿数字分页制作、手稿结构制作、元数据录入、质量监控、发布以及归还纸质手稿等步骤，如图 1 所示。图 1 左侧是手稿数字化的流程，右侧是对应人员在 Goobi-workflow 分布式系统中的工作流界面，每个人依据职位不同，登录系统后的工作流界面也不相同。这些步骤中需要的工具，都可以通过 Goobi-workflow 提供的插件完成，比如质量检查插件、图像上传插件等。Goobi-workflow 除了将复杂的手稿扫描分解为简单步骤，还具有强大的人员管理功能，以任务的形式，将分布在全球任何地方的人员组织起来，每个人通过 Goobi-workflow 提供的账户来管理项目管理员分配给的任务，而且还可以

图 1　华东师范大学图书馆 Goobi-workflow 工作流

图 2　华东师范大学手稿数据库

多个人共同完成同一类任务。如果对任务有疑问，还可以在任务中提问，并得到相关人员的解答。数字化完成的手稿，可以进行导出，或者直接发布到数字对象在线访问程序 Goobi-viewer，Goobi-workflow 可以为多个 Goobi-viewer 提供数字对象，只需要在创建项目时，设置好数字化资源保存的路径为对应 Goobi-viewer 的读取路径就可以。

3.2.2　图书分页与结构元数据

对于包含多页的数字对象，它的元数据不仅包括如题名、作者、摘要与关键字等元数据，还可添加分页与结构元数据，以此给用户提供良好的在线浏览体验。例如，图书书稿中的前言、引言、目录、附录等内容，常常与图书的主题内容采用不同的页码编辑，如前言使用罗马数字，正文使用阿拉伯数字，对于这种情况，通过对扫描的图书进行分页，不同分页的图像单独赋予编号。而对于图书的主体内容，会包含树状的章节结构，甚至注释信息，这时可以赋予图书结构元数据。

3.2.3　发布

使用 Goobi-workflow 创建的数字化手稿资源，可以直接使用 Goobi-viewer 进行在线发布，华东师范大学手稿数据库的发布界面如图 2 所示，左侧是资源分类浏览，右侧是数字对象浏览。Goobi-viewer 是一个可定制化的数字对象在线发布程序，Web 界面可以自适应电脑、平板以及手机等多种访问终端，图书馆可根据需要设置站点布局、菜单、页面功能等，对于有版权争议的资源，还可提供基于用户组、IP 等进行访问控制功能。内置 IIIF 图像接口与展示接口，默认数字对象使用 Goobi 自主开发的 IIIF 浏览器打开，同时也提供 Mirador、DFG-Viewer 浏览工具访问数字对象。除了提供手稿的在线浏览外，还提供 METS、MARC XML、Dublin Core 、RIS 等格式的元数据下载，以及 IIIF manifest、PDF 格式的数字对象共享，手稿数据库为每一个数字对象提供了唯一的 ID，同时动态生成 APA、MLA、Chicago 等格式的引文。

在资源发现方面，Goobi-viewer 具有强大的时空浏览功能、标签云以及全文检索等资源发现功能，检索结果提供缩略图、列表、详细信息等多种浏览方法，还可以根据时间、作者、地点进行排序，也可以将检索结果生成 RSS 链接，并保存到本地的 RSS 阅读器中。Goobi-viewer 并不完全依赖 Goobi-workflow 提供数字对象，它本身可上传本地数字对象，嵌入 IIIF manifest、OAI 收割等功能。

用户可以注册为 Goobi-viewer 手稿系统的会员，从而保存自己感兴趣的手稿，甚至可以保存搜索结果，并将这些记录统一发送到自己的邮箱。成为手稿库会员之后，用户还可以对手稿进行注释与转录等操作。

3.3　整合外部资源

通过对剑桥大学、哈佛大学、牛津大学、芝加哥大学等世界知名高校数字图书馆馆藏的调查发现，手稿是这些数字图书馆主要的资源类型之一，并且这些资源绝大多数是基于 IIIF 发布，IIIF 是数据层开放的数字对象发布方式，Goobi-viewer 数字对象发布工具内置多种 IIIF 浏览工具，能够轻松将这些开放的手稿资源整合到本地数据库。当然，也可以利用 Goobi-workflow 将整理国外开放手稿资源以任务的形式进行分发，基于团队形式进行更加系统的建设。收集到的开放手稿，也可以将数字对象本身缓存在本地，进行长期保存，以提高读者的访问速度。本研究结合所在高校的学科设置，对已经收集到的基于 IIIF 发布的部分手稿进行了简单的分类，如表 1 所示，这些开放的手稿将被全部整合到本地系统中。

3.4　针对手稿的图像浏览功能

为了提高某些字迹模糊手稿的在线阅读体验，基于 Goobi-viewer 创建的手稿数据库，改变图像对比度、亮度、

表 1　部分世界知名机构发布的手稿资源

	名称	数量	所属机构		名称	数量	所属机构
名人手稿	达尔文手稿	240	剑桥大学	古代语言手稿	埃塞俄比亚语手稿	16	剑桥大学
	达尔文－胡克信件	1168	剑桥大学		希伯来语手稿	18	剑桥大学
	牛顿手稿	48	剑桥大学		伊斯兰语手稿	88	剑桥大学
	莫扎特手稿	100	哈佛大学		梵文手稿	1600	剑桥大学
	肖邦早期手稿	400	芝加哥大学		英国中世纪希腊语手稿	425	剑桥大学
	林肯手稿	51	芝加哥大学		纳西族手稿	601	哈佛大学
乐谱	英国早期音乐手稿	24	剑桥大学		伊斯兰手稿	813	哈佛大学
	中世纪音乐	84	牛津大学		古代罗马手稿	390	罗马国家中央
绘画手稿	中世纪和西欧文艺复兴手稿	250	哈佛大学		古代中文手稿	514	牛津大学
	东亚画卷的摹本	300	哈佛大学		希腊语和拉丁语纸莎草纸	512	牛津大学
	伊斯兰石版画收藏	330	芝加哥大学		阿拉伯语手稿	130	牛津大学
科学	西方中世纪手稿	3095	牛津大学		亚美尼亚手稿	94	牛津大学
	英国中世纪医学手稿	40	剑桥大学		波斯语手稿（1602）	74	牛津大学
政治法律宗教	英国 1500 至 1800 年手稿	272	牛津大学		中世纪希伯来语手稿	832	牛津大学
	英国从 1801 年至今的手稿	706	牛津大学		中世纪拉丁语手稿	2015	瑞士虚拟手稿馆
	格里高利圣咏手稿	147	芝加哥大学		中世纪希伯来语手稿	158	瑞士虚拟手稿馆
	早期美国神学	100	耶鲁大学		中世纪德语手稿	788	瑞士虚拟手稿馆
	奴隶制和契约奴役收藏	39	芝加哥大学		中世纪法语手稿	243	瑞士虚拟手稿馆

锐化、色相等功能，提高模糊手稿的清晰度，如图 2（右）所示。如果要查看手稿中某一部分的细节，基于 IIIF 发布的手稿还可进行深度缩放。

3.5　注释转录工具

Goobi-viewer 默认的图像浏览工具虽然图像解析效率高，但仅提供整个图像的转录，无法对图像中的某一部分进行标记，为此本研究借助 Mirador 图像浏览工具进行手稿标注，它还可以实现手稿的对比阅读，如图 3 所示，左侧窗口是保存于牛津大学博德利数字图书馆中的莎士比亚手稿 [28]，右侧窗口是保存于本地手稿库中的高莽撰写《莎士比亚传》时的手稿，这种对比阅读能够激发读者更多的思想火花。手稿注释基于 Mirador 的 Annotations 插件，图 3 的右侧窗口，可以使用多边形、圆形、折线或者曲线选中手稿中需要标记的内容，然后进行注释，这里还可以修改标记图形的颜色，注释的字体与颜色等，一张手稿图像中可以添加任意多个注释，注释列表会在窗口的左侧显示。

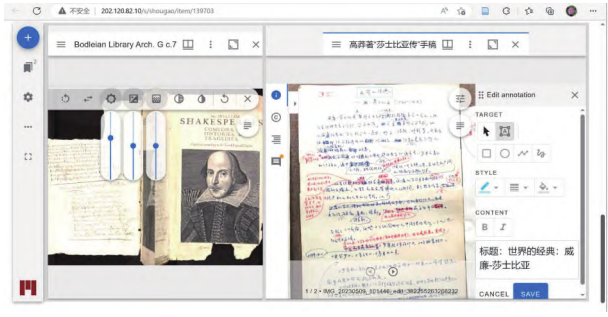

图 3　手稿对比阅读、转录与注释

3.6　人工智能技术的应用

3.6.1　手写识别

识别手稿中的文字，对于构建手稿全文检索、自动添加标注、融入大语言模型进行运算都非常有意义。然而当前以图书馆自身的技术力量对海量的手稿进行文本识别还是比较困难，为此，结合手稿数字化工作流中的 OCR 步骤，为手稿 OCR 人员提供人工智能工具，能够提高其工作效率。本研究调研了百度人工智能、阿里云、腾讯人工智能等国内人工智能技术公司，发现他们都具备手写识别开放平台，调用这些接口对于识别图书馆收集的部分手稿具有实际意义，能够完成部分手稿的转录工作，在一定程度上实现全文检索。以华东师范大学王元化《莎士比亚评论译丛》手稿为例，在百度、阿里、腾讯等平台中测试，发现百度的手写识别非常准确，甚至超过了普通的专业人员 [29]，识别结果如图 4 所示。

3.6.2　对象识别并提供聚类浏览

手稿除了文字本身，还包含印章、字形、纸张等信息，手稿数据库通过调用百度的通用图像与印章识别接口 [30-31]，识别出手稿中是否包含印章，手稿中撰写的字形是钢笔字还是书法等，这些信息将以标签的形式标注在每一张手稿中，成为读者进行资源发现的线索。百度的印章识别接口会返回印章的坐标信息，将其传输给 IIIF 的图像接口，就能批量将整个数据库中的所有印章全部抽取出来，以胡焕庸 1989 年的《邑人先辈童斐伯章先生亲书宜兴小志题跋》手稿中的印章为例，百度印章接口返回的印章坐标为“1881,2353,2051,2529”，在手稿数据库中其印章的 IIIF 网址为：http://example/*/1881,2353,2051,2529/pct:25/0/1989 年

图 4　人工智能对王元化手稿的识别与分析

胡焕庸撰《邑人先辈童斐伯章先生亲书宜兴小志题跋》.jpg，其中“pct:25”是需要获取印章图像的尺寸，“0”表示不旋转，这样就可以将整个手稿数据库中的全部印章图像单独整理出来，基于印章的作者进行分类，为读者提供印章图像集成访问功能。

3.6.3　使用 ChatGPT 类大语言模型

在图 2 中，华东师范大学手稿数据库已经对王元化《莎士比亚评论译丛》进行了手写识别，获取的文字信息中包含大量的实体，调用类似 ChatGPT 的大语言模型接口能够轻松获取这些文本中的实体信息，比如将《莎士比亚评论译丛》的文本提交给 ChatGPT，并使用提示词：“您只能使用提供给您的文本进行回答。您无需编造额外信息，答案必须包含在提供给您的文本中。从下面一段文字中提取的实体，请返回 JSON 编码”，命令 ChatGPT 进行分析，最后 ChatGPT 返回了如图 4 的 JSON 代码。

ChatGPT 给出的 JSON 格式实体记录，可以直接嵌入到手稿数据库的标签字段中，实现手稿资源分类或者聚类。随着手稿数据库使用人数的增多，对于无法机器手写识别的手稿，通过众包方式人工转录，能够不断进行大语言模型分析，增加每一篇手稿的标签记录，更便于读者使用。

4　思考与展望

随着人工智能技术的不断迭代，手写智能识别、图像自动标注、生成式大语言模型的不断发展，未来将能够更多地挖掘出手稿数据库中隐藏的知识，还能够实现辅助手稿鉴定等功能。

4.1　文本化与标注的不断深入

手稿识别技术的进步结合师生参与识别结果的纠正，将会使手稿资源不断文本化，大量的文本使全文检索得以实现。大量的文本也方便机器对手稿中的实体进行识别，为手稿创建更加丰富的标签，当然师生参与手稿众包也能产生标签，对这些标签进行聚类发布，聚类的标签将成为读者浏览手稿的线索。

4.2　基于角色扮演的智能咨询服务

大语言生成式聊天机器人的智力涌现，已经使其具备角色扮演的能力，而手稿数据库中不同人物的手稿数据，将成为角色扮演机器人最好的数据源。比如将华东师范大学手稿库中王元化先生的笔记、书信等资源，注入到 ChatGPT 并进行训练微调，就能实现 ChatGPT 扮演王元化先生的功能。当然还可以扮演吕思勉、胡焕庸等学者，对读者的提问进行回答，在生成式聊天机器人智力涌现的环境下，这样的角色扮演式咨询服务，能够极大地发挥出手稿数据库的特殊价值。

4.3　手稿鉴定

图书馆通过捐赠、拍卖等渠道获得了大量的手稿，手稿鉴定却成了一项艰巨的任务，当前的做法需要手稿专家来进行一一甄别，需要大量的人力和时间成本，小范围的甄别尚可通过这种方式实现，对于大量手稿图书馆将无能为力[32]。本研究探索机器学习技术在手稿鉴定方面的应用，在机器学习的帮助下提取手稿库中每个人物的手稿特征并保存在手稿数据库中，当有新的手稿需要鉴定时，将新手稿的图像上传到手稿数据库，系统就会自动给出鉴定结果。

手稿特征值不仅可以用于手稿的鉴定，还可以模拟他人的手稿笔迹，生成新的内容，这一点对于高校的文脉传承有着重要的现实意义。比如华东师范大学可以模

拟吕思勉、王元化等名师的笔记，来撰写新生的录取通知书。

4.4 多光谱三维手稿

当前基于二维扫描的数字化手稿会损失掉手稿中的许多细节，而这些细节往往是历史学家、社会学家非常需要的。本研究已经在探索构建三维数字手稿，当前的做法是采用激光雷达构建手稿的三维模型，然后再获取手稿的视觉信息，第三步再使用红外线、紫外线获取人眼无法直接察觉的色彩细节，从而恢复手稿中由于磨损或者泡水造成的损坏。为了将多光谱三维数字手稿呈现给读者，除了捕获手稿的三维模型，还需要开发可在线访问的三维图像浏览工具，并提供不同光谱的可视化效果，当前还没有合适的开源软件，需要自主开发。未来这项技术将首先应用于古老珍贵的手稿。

5 结语

手稿特藏具有重要的历史、文化价值，对支持高校的教学科研有着不可替代的重要作用，而且知名高校大都重视纸质手稿的收藏，但由于缺少手稿资源数字化、长期保存以及在线发布的解决方案，导致手稿只能保存于图书馆的书架，无法被师生使用，无法发挥其巨大的价值。本研究在分析了高校图书馆、师生、高校对手稿数据库不同需求的基础上，以华东师范大学图书馆手稿数据库建设为例，介绍利用 Goobi 开源软件，构建分布式手稿资源数字化加工与集成发布相结合的整体解决方案，构建符合 IIIF 图像规范与展示规范的数字对象互操作方案，使得本地手稿数据库具备集成剑桥大学、芝加哥大学、哈佛大学等机构数字手稿的能力，且已经梳理嵌入了部分开放的手稿。手稿数据库还考虑到读者阅读手稿的特殊需求，构建手稿对比阅读、图片调整工具、注释转录功能。为了更加深入地揭示手稿中的隐藏知识，手稿数据库可调用百度手写识别、图像识别接口，对手稿数据库进行内容标注与全文识别，并将识别的印章坐标与 IIIF 技术结合，构建独立的印章图片集成浏览工具，识别的手写文本与 ChatGPT 结合，自动抽取文本中的实体，并返回 JSON 格式的实体记录，作为手稿的标注信息。

随着手稿数据库收藏的不断增加，本研究还将探索手稿智能鉴定、笔迹模拟等功能。为了让师生能够更加深入地研究手稿，还将探索构建多光谱三维手稿在线发布功能，为师生提供身临其境的手稿研究体验。本研究的手稿数据库平台是对手稿资源分布式数字化加工、长期保存以及在线揭示的一次探索，希望能够推动国内图书馆手稿数字化的发展。

（来稿时间：2023 年 11 月）

参考文献：

1. 程焕文，黄梦琪 . 在“纸张崇拜”与“数字拥戴”之间——高校图书馆信息资源建设的困境与出路 [J]. 图书馆论坛，2015，35（4）：1-8.

2. 陈思和 . 试论高校图书馆特藏建设的意义 [J]. 杭州师范大学学报（社会科学版），2020，42（1）：1-6.

3. University of Cambridge.Cambridge Digital Library[EB/OL].[2023-03-30].https://cudl.lib.cam.ac.uk.

4. Harvard-Yenching Library.Naxi Manuscripts[EB/OL].[2023-04-12].https://library.harvard.edu/collections/naxi-manuscripts.

5. The University of Chicago Library.Collections & Exhibits[EB/OL].[2023-04-16].https://www.lib.uchicago.edu/collex/?digital=on&view=collections.

6. Yale University Library.Digital Collections[EB/OL].[2023-04-13].https://collections.library.yale.edu/catalog/2002046.

7. 复旦大学图书馆 . 卿云书房 [EB/OL].[2023-05-07].http://www.library.fudan.edu.cn/qysf/list.htm.

8. 上海交通大学 . 李政道数字资源中心 [EB/OL].[2023-05-21].https://tdlee.sjtu.edu.cn.

9. 北京大学图书馆 . 纸质文献数字化服务 [EB/OL].[2023-05-22].https://www.lib.pku.edu.cn/portal/cn/xxzc/szjg.

10. University of Cambridge.Darwin Manuscripts[EB/OL].[2023-05-26].https://cudl.lib.cam.ac.uk/collections/darwin_mss.

11. Bayerische Staatsbibliothek.Manuscripts personal papers and autographs[EB/OL].[2023-04-14].https://www.bsb-muenchen.de/en/collections/manuscripts.

12. 程焕文，张琦，谢小燕，等 . 向人类历史上最伟大的思想家致敬——中山大学图书馆藏马克思恩格斯手稿与珍稀著作的时代价值 [J]. 中国图书馆学报，2021，47（4）：4-15.

13. 陈思航 . 基于手稿资源的特色数据库建设 [J]. 图书馆工作与研究，2017，255（5）：51-56.

14. 刘荣弟，景月亲，张伦敦 . 探索数字时代音乐文献服务的新路径——西安音乐学院“特色数据库”建设与展望 [J]. 交响（西安音乐学院学报），2017，36（3）：136-141.

15. 陈以敏，张青青 . 数字人文下高校图书馆手稿特色数据资源库建设研究 [J]. 图书馆，2021，321（6）：87-93.

16. 张慧丽 . 康奈尔大学图书馆珍稀资源与手稿部服务及启示 [J]. 图书馆理论与实践，2014，181（11）：105-109.

17. Paucar-León V J, Molina-Granja F, Lozada-Yánez R, et al. Model of Long-Term Preservation of Digital Documents in Institutes of

Higher Education[C].Knowledge Management in Organisations: 16th International Conference, KMO 2022, Hagen, Germany, July 11–14, 2022, Proceedings. Cham: Springer International Publishing, 2022: 257–269.

18. 陈涛，张靖，赵宇翔，等 . 数字人文实践中特藏资源的关联数据实现机制探索——以方志资源为例 [J]. 情报理论与实践，2022，45（7）:180–187，147.

19. Matt Miller.Using GPT on Library Collections[EB/OL].[2023–03–30].https://thisismattmiller.com/post/using–gpt–on–library–collections.

20. Kota S S, Massand R, Agrawal A, et al. Digital Enhancement of Indian Manuscript, Yashodar Charitra[J]. Jaipur, India: Computer Science and Engineering Department, The LNM Institute of Information Technology, 2014.

21. 何光伦 . 名人手稿的典藏、保护与利用刍议——以四川省图书馆馆藏刘咸炘手稿为例 [J]. 图书馆杂志，2018，37（12）: 69–73.

22. 华东师大出版社 . 华东师大出新书贺杨振宁百岁寿辰 [EB/OL].[2021–09–24].https://www.ecnu.edu.cn/info/1094/58005.htm.

23. Hoover Institution or Stanford University.Chiang Kai–shek & Chiang Ching–kuo Diaries[EB/OL].[2023–05–25].https://www.hoover.org/library–archives/collections/featured/chiang–diaries.

24. 上海交通大学出版社 . 中国地方历史文献数据库 [EB/OL].[2023–05–28].http://ndfwx.datahistory.cn.

25. 澎湃新闻 . 来上图手稿主题馆看展：“读书这么好的事”[EB/OL].[2023–03–30].https://www.thepaper.cn/newsDetail_forward_22832345.

26. 胡晓明，周兴陆，刘锋杰，等 . 王元化与后五四反思（笔谈）[J]. 华东师范大学学报（哲学社会科学版），2019，51（4）:1–16，185.

27. 华东师大 . 华东师大学人手稿文献展开幕 [EB/OL].[2023–04–24].https://www.ecnu.edu.cn/info/1094/61456.htm.

28. Bodleian Librariy.Shakespeare's Dead exhibit [EB/OL].[2023–06–17].https://blogs.bodleian.ox.ac.uk/digital/2016/04/23/introducing–the–iiif–first–folio.

29. 百度大脑 . 手写文字识别 [EB/OL].[2023–04–23].https://ai.baidu.com/tech/ocr_others/handwriting.

30. 百度大脑 . 印章识别 [EB/OL].[2023–04–27].https://ai.baidu.com/tech/ocr/seal.

31. 百度大脑 . 通用物体和场景识别 [EB/OL].[2023–05–17].https://ai.baidu.com/tech/imagerecognition/general.

32. 王金坪 . 朱万章：批量画稿需要逐张鉴定 [J]. 收藏，2016，329（21）:142–143.

Practical Research on Digitization and Online Publishing of Manuscripts in University Libraries: Case Study of Manuscript Database Construction in East China Normal University Library

Zhang Yi

(East China Normal University Library)

【Abstract】 Domestic university libraries have collected a large number of paper manuscripts, but due to the lack of overall solutions for digitization, long-term preservation and online disclosure, their value cannot be effectively utilized. Based on the analysis of manuscript database requirements, this study takes the construction of manuscript database of East China Normal University as an example, uses the open source software Goobi to realize distributed manuscript processing and publishing, and uses artificial intelligence technology to dig deep into manuscripts. The underlying architecture based on IIIF enables the special collection to integrate manuscript resources published by internationally renowned cultural heritage institutions to the local. According to the characteristics of manuscript resources, build manuscript online reading image tools, annotation transcription, handwriting recognition, feature recognition, automatic labeling and other functions to improve user experience. In the future, we will continue to explore functions such as celebrity role-playing reference consultation, multi-spectral 3D manuscript browsing, and manuscript identification to promote the construction of manuscript databases in domestic university libraries.

【Keywords】 University library　Manuscript database　Goobi　Distributed workflow　ChatGPT

〔作者简介〕 张毅（1986—），男，研究生，华东师范大学图书馆副研究馆员，研究方向：数字人文与数字特藏资源建设。