

智慧数据驱动的古籍智慧性保护体系研究

陈 涛 苏日娜 张永娟 张 靖 余厚强

摘 要 古籍承载着中华民族的共同记忆,对古籍进行保护、数字化、活化利用,就是延续中华民族的文化命脉。近几年,智慧数据的兴起给古籍保护提供了新的研究和实践视角。本文梳理总结现有的智慧数据研究成果,指出智慧数据是一个定性的概念,是一个流变的过程,是一个多态的结果;提出智慧数据产生的技术路径,将其划分为语义数据、关联数据、智能数据和智慧数据四大数据形态;在智慧数据的驱动下,提出古籍智慧性保护体系框架,从数据的角度、知识的维度和智慧的深度整体构建古籍知识谱系。智慧性保护丰富了古籍数字化的研究范畴和保护方法,贯彻了文化遗产保护发展理念,推动了古籍活化利用的实现进程,并迎合了科技发展的趋势。正如智慧没有边界,智慧性保护的内容和范畴也不会固定,会随着技术的进步而不断得到完善,未来智慧性保护将会得到更多的认可和关注。图3。表1。参考文献28。

关键词 古籍保护 智慧性保护 智慧数据 文化遗产 数字人文

分类号 G255.1

The Smart Conservation System of Ancient Books Driven by Smart Data

CHEN Tao, SU Rina, ZHANG Yongjuan, ZHANG Jing & YU Houqiang

ABSTRACT

The preservation, digitization, and revitalization of ancient books, which carry the common memory of the Chinese nation, is to continue the cultural lifeline of the Chinese nation. In recent years, the rise of smart data has provided a new research perspective for the digitization and conservation of ancient books. This paper points out that smart data is a qualitative concept, a rheological process, and a polymorphic result. First of all, it is difficult to have unified quantitative indicators for smart data, and it needs to have suitable scenarios and specific periods to reflect the value of data. Secondly, smart data is not static, and with the increase of applications, the increase of associations, and the enhancement of content, the value of smart data will also show dynamic characteristics. Finally, smart data does not only contain a certain data form, but forms a knowledge community by multiple data forms in the process of dynamic development.

On the basis of smart data, this paper proposes the concept and framework system of smart conservation of ancient books. The smart conservation takes smart data as the main form of expression, and the implementation process is divided into semantic data, linked data, intelligent data, and smart data. Among them, semantic data can be regarded as the primary form of smart data; intelligent data is the lightweight smart data, which favors data dynamics and automation; and smart data is the enhanced version of

通信作者:余厚强,Email:yuhq8@mail.sysu.edu.cn,ORCID:0000-0002-9241-6630 (Correspondence should be addressed to YU Houqiang,Email:yuhq8@mail.sysu.edu.cn,ORCID:0000-0002-9241-6630)

intelligent data, which is more inclined to reproduce knowledge.

Subsequently, the paper points out that smart conservation is an intermediate form between regenerative conservation and heritage conservation of ancient books, which provides a smarter and more intelligent knowledge source for heritage conservation of ancient books. Smart conservation requires the overall construction of the knowledge spectrum of ancient books, which will contain multimodal data (such as original ancient books, images, text, audio, and video), knowledge forms (such as metadata, ontology, entities, relationships, RDF data), semantic enhancement contents (such as entity links, annotations, knowledge), and even intelligent data such as anticipation, rules, AR/VR, knowledge graph, as well as algorithms, codes, models, etc., on which the smart generation depends.

Finally, this paper points out that smart conservation not only enriches the research scope and conservation methods of ancient books digitization, but also implements the development concept of cultural heritage conservation. Further, it may even promote the realization process of revitalization and utilization of ancient books. As the smart technology advances, the content and scope of smart conservation will not be fixed, and will continue to be improved. We hope that smart conservation can receive more recognition and attention.

3 figs. 1 tab. 28 refs.

KEY WORDS

Ancient books preservation and conservation. Smart protection. Smart data. Cultural heritage. Digital humanities.

0 引言

文化遗产是人类世代代的创造和积累,积淀着各个历史时期的杰出贡献,可以再现昨天、前朝甚至远古的历史风貌^[1]。古籍是中华民族创造的重要文明成果,是中华文明绵延数千年、一脉相承的历史见证,也是不可再生的文化遗产。古籍是中华优秀传统文化的重要载体,是前人留给我们的宝贵财富,我们有责任将其保护好、传承好、发展好。习近平总书记强调:“让收藏在博物馆里的文物、陈列在广阔大地上的遗产、书写在古籍里的文字都活起来。”“活起来”为文化遗产保护工作指明了方向,也只有保护好,才能实现“活起来”。

古籍保护有原生性保护、再生性保护、传承性保护三种类型。原生性保护就是保护古籍文物本身;再生性保护是指对古籍进行影印、数字化,并进一步整理;传承性保护,是指让古籍里的知识、故事、思想等进入人们的大脑,通

过这一载体传承下去^[2]。早在2007年,国务院办公厅印发《关于进一步加强古籍保护工作的意见》,正式启动实施“中华古籍保护计划”,该计划主要针对古籍原生性保护和再生性保护开展,古保计划推行十五年来,培养了一批古籍修复人才,大量珍贵古籍得到抢救性保护和修复。2017年《“十三五”时期全国古籍保护工作规划》、2022年《政府工作报告》等文件多次强调加强古籍资源保护利用。2022年4月25日,习近平总书记在中国人民大学考察时也强调“要运用现代科技手段加强古籍典藏的保护修复和综合利用”。2022年两办印发的《关于推进新时代古籍工作的意见》和《关于推进实施国家文化数字化战略的意见》,对古籍资源及优秀传统文化相关工作提出了更高的要求,如加强古籍资源管理和开放共享、关联形成中华文化数据库、共建文化数据服务平台、建设文化计算体系等。这些任务和要求俨然超出了简单的古籍利用的范畴,不仅要创建古籍资源的影印、数字副本、数据库等,更要共享古籍数据资源、汇聚

文化数据信息、关联不同形态(文字、音频、视频等)的文化资源数据,甚至要求将凝结文化工作者智慧和知识的关联数据转化为可溯源、可量化、可交易的资产。由此可见,需要对古籍原生性保护和再生性保护的内容进行拓展和延伸,以顺应时代需求。古籍保护不应停留于某一种或某几种保护类型,应随时代的发展、科技的进步和理念的创新逐渐衍生出新的保护理论和体系。多形态数据资源、文化计算体系、智慧和知识相关的资产等都应与古籍本身一样得到足够的重视,唯有全面保护好、充分利用好这些知识和体系,才能更好地提升古籍利用效率、挖掘古籍的时代价值,这也是开展古籍智慧性保护研究的主要立足点。

1 研究和实践综述

文化遗产保护是一个时代性的,也是一个历史性的使命,文化遗产保护应从“抢救性保护”阶段进入“科学保护”的新阶段^[3]。作为文化遗产的重要组成部分,文献遗产的保护框架和保护策略研究备受学者关注。保护框架方面,提出需要建立我国文献遗产目录、建立文献遗产保护数字工程^[4]以及构建文献遗产精准保护框架^[5];保护策略方面,提出建立跨部门的文献遗产保护事业共同体^[6],并加强国家交流合作,积极参与国际标准制定^[7]。

信息技术的发展给文化遗产保护带来了新的视野,信息技术在文化遗产存档、管理、信息共享等方面的优势得到普遍认可。文化遗产数字化已经成为文化遗产保护和新方向和新趋势,国际图联的文化遗产保护策略也强调对于数字化手段的运用^[8]。古籍数字化研究尤其重视使用数字化技术对古籍文本进行处理^[9-11],如数字化输入技术、智能化处理技术、GIS技术、古籍自动句读与标点、古籍词语自动切分、古籍命名实体识别、计算机图像识别等。随着人工智能技术各细分领域不断创新和发展,机器学习、智能计算、虚拟现实等也常应用

于古籍数据挖掘^[12]和古籍数字化再造^[13]。此外,古籍数字化平台建设也得到高度重视,如国家图书馆“中华古籍资源库”、中华书局“籍合网”、上海古籍出版社“汇典”等项目极大推动了古籍数字化和保护事业的发展。同时,学界和业界的联合也使古籍焕发全新生命力,如北京大学携手字节跳动推动中华古籍数字化平台建设,将通过OCR、句读、实体识别和知识图谱等技术构建一站式自动古籍智能化整理平台;阿里巴巴与四川大学、美国加州大学伯克利分校、中国国家图书馆、浙江图书馆跨界协作,发起“汉典重光”项目,致力于寻觅海外流散的中国古籍,推动其以数字化方式“回归”。

智慧数据作为文化遗产、古籍数字化和数字人文等领域的新兴研究方向之一,已得到学界的关注。2017年,武汉大学举办“面向数字人文的智慧数据建设专题研讨会”,与会专家一致认为智慧数据这一新兴概念是一面旗帜,将引领文化生产与记忆机构的资源组织、资源管理朝着更加先进、更加智能、更加智慧的方向发展。智慧数据是数据资源价值的承载^[14],是文化遗产资源价值挖掘、创造性转化和创新性发展的基础^[15]。智慧数据主要来源于大数据,从大数据中得出有意义的信息,以支持决策和行动^[16,17]。在智慧数据的具体实践中,语义技术和关联技术是常用的技术解决方案,如敦煌文化遗产智慧数据借助元数据、主题词表、本体、关联数据、语义增强、知识图谱等技术得以创建并提供服务^[18];地方历史文献智慧数据的构建使用自建元数据方案和实现资源数据的多重关联方法^[19];文化遗产档案智慧数据的资源建设遵循数据建模、数据关联、数据呈现的技术路径^[20]。2021年,武汉大学文化遗产智能计算实验室入选教育部哲学社会科学实验室,以及国家社会科学基金重大项目“文化遗产智慧数据资源建设与服务研究”立项,都将进一步推动智慧数据在文化遗产中的研究与发展。

通过上述分析可知,古籍的数字化研究和利用已成燎原之势,古籍保护体系也从单纯地

探讨保护途径转向更多地关注古籍科学保护与合理利用。基于智慧数据理论和相关技术,可以在古籍资源原有类型之外,构建更为丰富、更为智能、价值更高的知识体系,以秉承动态保护、科学保护、整体保护的古籍资源新型保护理念。

2 智慧数据内涵及特征分析

智慧数据源自对大数据价值的认可和挖掘,是为了实现大数据的价值而提出的,是大数据技术的进一步发展。虽然目前智慧数据尚无统一的定义,但对其的理解在很多角度具有一致性,比如智慧数据从大数据中产生,会带来数据价值的提升;智慧数据具有可解释特性,具有语义表示能力;智慧数据是为了更好地辅助决策等。

从智慧数据的内涵和特征来看,智慧数据具有自描述、人机可读、可解释、可溯源等特性^[18,21],在此基础上可实现智慧数据的智能决策和行动(可行动性)^[22]。智慧数据的价值取向呈现出多元化复合的特征,它的技术方案也并非固化,而具有不断动态更新的特点^[23]。除此之外,本文认为智慧数据是一个动态的变量,而非静态的常量。对于智慧数据的理解,至少可从定性评价、流变发展和多态组成三个特点来把握。

(1)智慧数据是一个定性的概念。“智慧”这个词很难有统一且量化的评价标准,很难简单地评判哪些数据的智慧程度高,哪些数据的智慧程度低。智慧数据更多的是从一种价值体现的角度来提出的,只有在适合的应用场景,甚至有些只有在特定的时期,才能体现出某些数据特有的价值。

(2)智慧数据是一个流变的过程。智慧数据并非静止不动,而应是一个动态流变的过程。某些数据也许近阶段并无智慧或仅有极少智慧,但随着应用的增多、关联的增加、内容的增强,将逐渐体现出其智慧的一面。同时随着数

字技术的发展和智能算法的不断演进,智慧数据的价值也将逐渐得到提升。

(3)智慧数据是一个多态的结果。智慧数据并非只含有某一种数据形态,如文本、影像或模型,而是在动态发展过程中,由多种数据形态综合形成的结果。多种数据形态之间相辅相成、互为补充,形成知识共同体,通过提供整体的决策来指导行动。相比单一的结果,智慧数据更应注重智慧的过程和知识的积累。

智慧数据上述三个特点之间并非完全独立,需从整体的角度来理解。正由于智慧数据定性的特点,因此不能用一个固定的量化标准和评估值,而应将其放置于一定的时间范围内进行衡量,这就形成了流变的前提;同时,流变也使得可以从更多的维度来开展数据研究,使得智慧数据形态多变(多态),多态的结果也加剧了智慧数据评判方法和指标的不确定性。

3 智慧性保护体系框架研究

在古籍数字化研究和智慧数据特征分析的基础上,本文试图从保护路径和技术路线两个维度构建古籍资源智慧性保护体系框架(见图1)。保护路径作为横向维度,主要为现有的古籍保护方法(原生性保护、再生性保护、传承性保护)和本文提出的智慧性保护。纵向维度为智慧数据的详细技术路线,涉及从原生数据到智慧数据的逐级演变过程,共分为数字化、文本化、概念化、关联化、智能化和智慧化六个步骤。数字化实现了古籍资源原生性保护到再生性保护的转变,在进行文本化的基础上逐渐过渡到智慧化,每一次过渡都对应技术路径中的一层递进。如图1所示,智慧性保护包含了古籍资源从概念化到智慧化演进过程中的中间数据,这些数据理应得到重视和保护,以免造成知识的重复构建或遗忘。

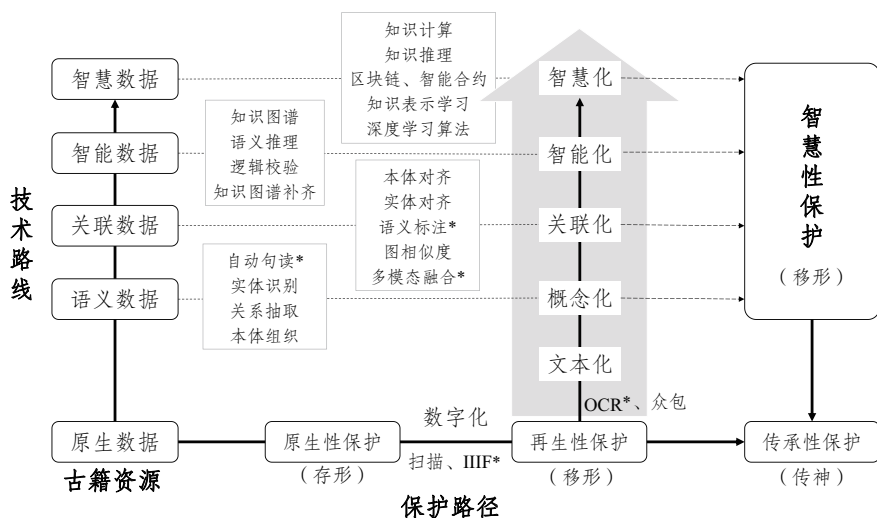


图1 古籍资源智慧性保护体系框架

整体框架在设计时考虑了古籍资源和内容的独特性,图中加星号(*)的方法是古籍资源智慧数据生成过程中最为值得关注的核心技术,主要体现在以下六个方面。①古籍年代久远,大量古籍残损严重,亟待深入推进古籍再生性保护研究,因此古籍智慧性保护将在古籍数字化基础上对古籍进行深入研究;②古籍资源离散,由于各种原因散落于海内外众多公私收藏机构、个人及古籍拍卖、经营市场,如何使用数字技术(如 IIIF)实现离散资源的数字化回归是古籍保护和利用的重要基础;③古籍版式多样,字体、行款、印章、批注、圈点、插图等给古籍文本 OCR 识别提出了挑战;④古籍字义丰富,尤其是生僻字、异体字、避讳字、多音多义字等,给古籍文本的理解、自动句读、实体抽取等带来很大难度;⑤古籍版本多杂,有原刻本、翻刻本、覆刻本,也有写本、稿本、抄本、刻本等不同版本,甚至包括文本、图像、音韵,如何将不同版本和数字形态汇聚融合,对统一古籍资源至关重要;⑥古籍批注校勘是前人在古籍研究和整理方面的知识积累,赋予古籍内容新的理解,使用语义标注方式将各类注疏、注释、注解或不同版本内容上的变化进行呈现和关联,可以为古籍研究提供更多的参考和研究视角。

3.1 智慧数据技术路线

语义网、知识图谱、区块链、人工智能、大数据、5G 等一系列数字技术的快速涌现与发展,给了人们无限的想象空间,也为实现古籍知识体系保护提供了技术手段。智慧数据是本文所提出的古籍“智慧性保护”的核心理论基础,具体而言,其产生过程主要包括以下几个阶段。

(1) 语义态:原生数据到语义数据的转化

古籍原体所对应的数据称为“原生数据”,主要包括元数据信息和数字化的图像资源。这些数据要想成为智慧数据,首先需要实现从原生数据到语义数据的转化。语义数据主要指使用资源描述框架(RDF)进行描述的古籍数据,既包括元数据信息,也包括古籍中的具体内容信息。以 RDF 为基础的语义数据已广泛应用于数字人文诸多知识库的构建当中,并成为关联数据应用和知识图谱研究的基础数据(图)模型。

原生数据到语义数据的转化主要包括数字化、文本化和概念化三个环节。再生性保护中的数字化更多指的是古籍原件的图像资源,而国际图像互操作框架(IIIF)已成为文化记忆机构中图像资源组织的首选方案,世界上主要的文化记忆机构(如大英图书馆、Getty 博物馆、法

国国家图书馆、OCLC、Europeana 等)都已在使用 IIIF 框架。古籍数字化后的图像中蕴含的知识需要经过文本化的过程,方可进行进一步的深度挖掘与利用,这里图像文本化的常用方法有 OCR 和众包。有了文本化的古籍内容后,自然语言处理(NLP)及相关的机器学习算法便有了用武之地,如可进行古籍句读、实体识别、关系抽取等。抽取的实体和关系需要使用本体进行组织,并使用 RDF 框架进行编码,形成概念化的语义数据,这也成为智慧数据的最初形态。也就是说,数字人文视野下智慧数据的最基本形态为语义(RDF)数据。之所以将 RDF 数据称为语义数据,是因为 RDF 数据的组织使用本体结构,从而赋予机器可理解的语义(本体中的类和属性)。

(2) 关联态:语义数据到关联数据的转化

语义数据使用本体进行组织及使用 RDF 进行描述,符合智慧数据的自描述、可解释和可追溯的特点,这一阶段的语义数据仍然处于价值孤岛,缺乏与外界数据的交换与联系,从而制约着知识的扩容和衍生。关联数据在数字人文和文化遗产领域中的应用价值已被反复证实,已被众多的古籍、档案、手稿等文献知识库所采纳。尽管本体的设计倡导尽量复用已有的本体,但复用并非易事,因而更多的本体在设计时仍然采用自行设计的方式,这就导致了本体结构的差异性,也给资源之间的关联带来了挑战。不同资源之间的关联通常采用本体对齐和实体对齐,以及文本相似度和图相似度进行判断。这些对齐和相似度的方法可以借助机器学习的相关算法来完成,并辅以人工审校来提高关联的精准度,这些关联多是针对结构化的文本资源。

每种信息或数据的来源或形式可以称为一种模态,如文字、图像、音频、视频、3D 模型等,多模态亦即表示两种或两种以上的资源形式。使用机器学习的方法实现非结构化的资源多模态之间的关联难度较大,比如将语义数据关联到某些图像中的局部区域,这涉及计算机视觉的相关内容,如图像的自动分类、图像的自动化语

义描述、图像的目标识别等任务。这些任务不失为多模态关联中很好的研究方向,但目前业内尚无成熟的研究成果。对于音频、视频、3D 模型等模态的知识关联,难度则更大。也许将来会出现自动化的关联流程和高精度的关联算法,但目前对于多模态之间的关联更多的是采用人工处理的方式。对图像资源进行内容注释和语义标注,也是建立多模态知识之间关联关系的主要方法之一。

(3) 智能态:关联数据到智能数据的转化

关联数据追求的是古籍知识的开放、共享,以及与不同模态的古籍知识的交互和融合,但这些知识之间的关联更多的是以一种显性(已存在)的形式出现。相比关联数据,智能数据更强调通过自动化的流程和设置来提升数据的质量,以及对隐性知识进行推理和挖掘。关联数据使用 RDF 模型进行资源描述,又使用本体结构进行组织,因此基于本体的推理和基于逻辑规则的推理将会是智能态中语义推理的主要形式。大规模知识图谱的自动构建也是提升数据智能化的主要手段之一。从构建技术看,知识图谱经历了由人工构建到群体智慧构建再到自动获取构建的过程。兼顾效率和精度,目前知识图谱最合理的创建方式为半自动结合人工,而纯人工或全自动的构建方法较为少见。除了构建知识图谱外,知识图谱中常涉及的数据存储、图计算、检索算法、知识图谱补齐等方面的技术亦可划归智能态的技术体系中。

(4) 智慧态:智能数据到智慧数据的转化

智能态可看成是轻量级的智慧态,智慧态则是升级版的智能态。两者之间的主要区别在于智能态更多的是倾向于自动化批量操作,而智慧态更倾向于知识的计算和再生产。智能态和智慧态之间并非完全割裂,两者之间的很多理念、技术和方法可以重叠,如知识图谱与知识推理。智慧态中可运用的知识推理主要指知识图谱中基于图结构的推理、基于分布式表示学习的推理和基于神经网络的推理,以及知识计算、深度学习算法等。智能态和智慧态中都

含有推理,为了区分,将智能态中的推理称为语义推理,主要是本体结构和关联数据中的逻辑规则,将智慧态中的推理称为知识推理,主要是借助知识图谱和深度学习算法进行的深度推理。此外,区块链、智能合约作为智慧态中的核心技术,其意义不仅在于让数据“有迹可循”“有源可溯”,更是为了打造一个更公正、更透明、更可信的数据生态体系。“分布式存储”“智能合约”“哈希算法”等区块链中的核

心技术将赋予数据更高的智慧,同时,区块链的使用也将会解决关联数据中长期存在的数据可信度的问题。

为了更清晰地理解智慧数据的特征,现对智慧数据产生过程中的不同数据形态进行对比(见表1)。从表1中可知,智慧数据的产生是一种动态、进化的过程,每一个数据态都是循序渐进产生,并包含该数据态之前的所有形态,各数据态之间的主要区别体现在以下几个方面。

表1 智慧数据中的各数据态

类型名称	数据态	主要功能	核心技术	主要特性	保护体系
原生数据 Original Data	原生态	古籍元数据及内容呈现	著录、机器扫描……	易传播 可利用	再生性保护
语义数据 Semantic Data	语义态	古籍(显性)知识的语义重组(编码)	OCR、本体组织、语义网、自然语言处理、实体识别、关系抽取……	自描述 可解释 机器可读	智慧性保护
关联数据 Linked Data	关联态	古籍(显性)知识的多模态融合	关联数据、数据挖掘、知识融合、实体对齐、本体对齐、相似度算法……	易开放 广关联 强交互	智慧性保护
智能数据 Intelligent Data	智能态	古籍(隐性)知识的自动生产	语义推理、逻辑校验、知识图谱(构建)、知识图谱补齐……	可推理 可自动 可追溯	智慧性保护
智慧数据 Smart Data	智慧态	古籍(隐性)知识的再生产	知识图谱(算法、推理)、知识计算、知识表示、区块链、智能合约、深度学习算法……	可推理 可计算 可行动 可发现	智慧性保护

(1)主要功能区别。语义态和关联态都是针对古籍中的显性内容而言,语义态实现了古籍知识的语义重组,关联态旨在实现古籍知识的多模态融合;智能态和智慧态从古籍隐性知识的挖掘和生产出发,智能态是为了体现古籍知识生产的自动化过程,而智慧态专注于古籍知识的再生产。其中,语义态、关联态和智能态之间的区别较为明显,而智能态和智慧态之间的差异往往并不明显。

(2)主要特性区别。语义数据采用RDF和本体组织,因此具备了自描述、可解释和机器可读的特点;关联数据则是通过开放和关联,实现

不同模态数据之间的交互;智能数据中的语义推理、逻辑校验和知识图谱等技术确保了数据具有可推理、可自动和可追溯的特点;智慧数据同样具有可推理的特点,但更多的是具有可计算、可行动和可发现的能力。

(3)保护体系区别。原生数据中的元数据和扫描的数字化资源是为了古籍资源的再生性保护;从语义态到关联态、智能态和智慧态,都可以看成是智慧性保护的范畴。一定意义上讲,语义数据的产生是一种最低层级的智慧保护,智慧数据则是智慧保护的高级形态。

当然这里探讨的智慧也许不能等同于人的

“智慧”,近阶段可看作弱智慧。未来,科学技术无疑将会有更迅猛的发展,人们也渴望智能技术有更广泛的应用,越来越多具有人的“智慧”的机器、技术或算法的推出,将会不断拓展智慧数据的边界。

3.2 智慧性保护体系构建

长期以来,我国古籍保护从单纯的原生性保护,逐渐发展为“原生性+再生性+传承性”保护的立体格局,形成了涵盖古籍保存保护修复、古籍编目整理和研究出版以及古籍活化利用传播等全域全流程的系统性工作。①原生性保护是不改变古籍原始特征和载体形式的一种保护形式,包括古籍修复、加固及改善藏书环境等,主要是保护古籍实物(原体)。抢救性保护也属于原生性保护的范畴。②再生性保护是改变古籍原始特征和载体形式、对古籍内容进行复制的一种保护形式,保护手段包括影印、数字化、微缩复制、整理出版等。此处的数字化主要指对古籍原体的图像和文字等内容的提取。③传承性保护理念根植于非物质文化遗产领域,形式是活态传承,由传承主体、传承内容、传承场所、传承方法构成。其本质是古籍保护从业者和爱好者通过各种形式的宣传推广、展览讲座、展示体验、民俗演示、网络传播、文创开发等活动深入诠释、解读蕴藏在古籍中的优秀传统文化、学术和艺术。

原生性保护是保存古籍实物,可谓之“存形”;基于内容复制的再生性保护是“移形”;以增进古籍保护意识、传承优秀传统文化为宗旨的传承性保护是“传神”^[24]。其中再生性保护中的数字化也仅仅是对古籍内容的复制,更多的古籍图像形态的数字化并非完全基于古籍内容而言。然而,上述三层保护体系中,针对古籍内容的保护维度并不清晰和完备,或者更

进一步讲,保护体系中缺少针对知识内容层面的保护范畴。一代代传承人或研究者对于古籍知识的整理、加工、理解以及再生产的过程,同样需要进行记录、留存和保护,从而为后来人提供坚实的研究基础和知识来源,而这一点恰恰是目前古籍保护体系中缺失的重要一环,智慧性保护体系会重点关注并研究这部分内容。所以,智慧数据驱动的“智慧性保护”体系的关注对象不再仅是古籍的原始数据,而是包含了古籍内容价值提升过程中的各阶段数据,这些数据也将成为“移形”中的一部分,共同为“传神”提供更为丰富、更为全面、更为智慧的知识来源。

智慧性保护可以作为再生性保护和传承性保护的中间形态,各保护类型之间的关系如图2所示。智慧性保护极大丰富了再生性保护的范畴,将再生性保护的对象类型从数字化形态的图像、影印、胶卷等转换到数据甚至是知识层面。如果说再生性保护为传承性保护提供资源,智慧性保护则为古籍活化和研究利用等传承性保护提供更为智能、更为智慧的知识源。从图2中可以看出,智慧性保护的内容包含了智慧数据技术路径实现过程中各个阶段产生的中间数据。具体来看,可以将智慧性保护的范畴归纳为图3,整个过程需要根据实际情况依赖于一定的基础设施、研究平台和研究工具。其中,基础设施包括网络基础设施^[25]、数据基础设施^[26]、语义发布基础设施^[27],以及各种数据中台、图像中台、算法中台等;研究平台有中国历史地理信息平台^①、学术地图发布平台^②、古籍自动整理平台^③、Docusky 数位人文学术研究平台^④、历史人文大数据平台^⑤、多维度图像智慧系统^⑥,以及自主研发或者第三方商业公司的OCR识别平台,如书同文“i-慧眼”古籍汉字自动识

① <http://timespace-china.fudan.edu.cn>

③ <http://wyd.kvlab.org>

⑤ <https://dhc.library.sh.cn>

② <http://amap.zju.edu.cn>

④ <https://docusky.org.tw>

⑥ <http://miss.newwenke.com>

别、自动录入平台^①等;研究工具主要有本体开发工具(如 Protégé)、数据库工具(如 Neo4j)、数

据清洗工具(如 OpenRefine)、标记训练工具(如 Markus)、可视化工具(如 Gephi)等。

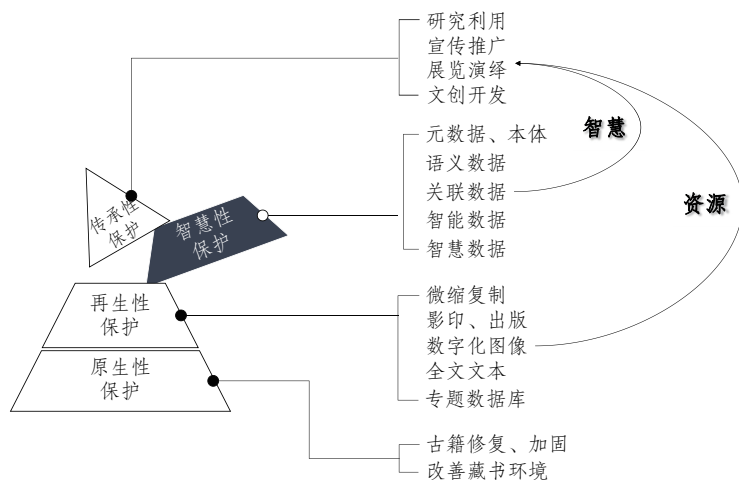


图2 古籍保护体系结构

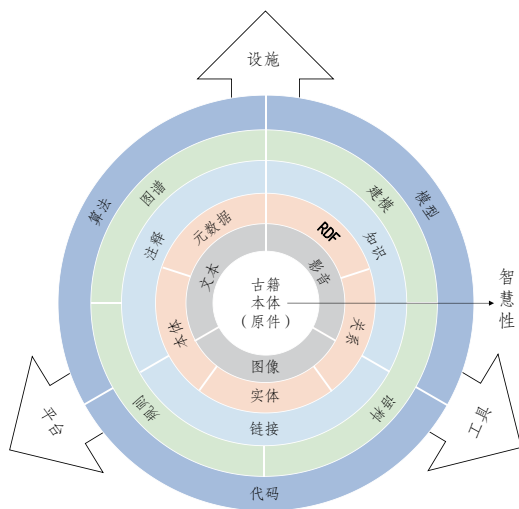


图3 智慧性保护的内容范畴

基础设施、研究平台和研究工具的共同作用,将给古籍数字化研究带来量变和质变。当然,为了弥补资金投入差别造成的学术鸿沟,可以加强计算工具平台、数据资源的共享^[28],以帮助偏远地区或者学术资源不足的地方开展研究。通过共享与合作,设施、平台和工具将筑成

古籍数字化基础设施,共同支撑古籍资源的保护、研究和利用。在古籍数字化基础上开展的智慧性保护的核心内容包括以下五项。①再生性保护将会在原生性保护的成果(古籍本体,即原件)基础上产生文本、图像和影音这三类主要数据。②智慧性保护过程中的语义数据将主要

① <https://dzcj.uni-han.com.cn>

产生元数据、本体、实体、关系以及映射的 RDF (三元组)等类型数据。③关联数据环节主要涉及注释、链接和知识三部分,注释包括文本注释、图像注释、音视频注释等。链接为古籍与其他外部资源以及知识库的关联地址;考虑到链接地址容易失效,可以将链接后的知识进行缓存,以实现持久保存。④智能数据环节需要记录生成的知识图谱、语义推理的规则、逻辑校验的规则、机器学习的语料、建立或关联的 AR/VR 等数字模型。⑤智慧数据阶段主要是对知识计算和智能计算的算法、应用或调优的模型,甚至是对所有过程中涉及的代码(程序)进行保存。当然,在后两项中同样需要对产生的知识进行保存和保护。

非物质文化遗产保护中强调整体性保护原则,整体性保护理念可以用来佐证智慧性保护方法的重要性。非物质文化遗产的整体性保护原则提出了从空间向度和时间向度上共同构建文化遗产的立体保护体系,强调非遗保护不能离开其赖以生存的环境,以及与之对应所形成的文化价值观念。古籍保护同样需要整体性保护原则,原生性保护中的古籍原体、再生性保护中的数字资源、智慧性保护中的知识体系以及传承性保护中的衍生形式,都是围绕古籍构建的内容整体。不同类型保护之间应互为依赖、互相补充,即再生性保护需要依赖原生性保护的保护质量,智慧性保护需要依赖于再生性保护中的数字资源完备程度,传承性保护中演绎的内涵、文创的理念也都需要依赖于对古籍内容的理解程度以及智慧性保护中的知识体系。基于智慧数据技术构建的智慧性保护除了具有整体性、智慧性外,还伴有重用性和传承性的特点,保护的知识体系也将随着知识的传播和文化的传承不断地动态传递下去,同时也在演变中逐渐得到完善和扩展,以便构建更为全面、更为智慧的知识谱系。

4 古籍智慧性保护的再思考

关于古籍保护的今后发展,无论是原生性、

再生性、传承性或是本文提出的智慧性保护,都需要遵循一套核心话语体系,且在一种可持续推动框架内进行。本文提出的智慧性保护框架即从古籍的内容及知识层面试图为此提供解决路径,以真正实现让“书写在古籍里的文字都活起来”。然而,上述关于古籍智慧性保护的理论体系面对的对象是洋洋大观的典籍文献,要真正做到活化利用、内容挖掘、知识诠释和智慧传递,详解古籍中所蕴藏的思想精华,仍然面临诸多困难。因此,本文重新审视古籍保护这一宏大主题,希望通过以下几个方面,进一步完善古籍资源智慧性保护的理论体系。

4.1 古籍智慧性保护丰富了古籍数字化研究范畴和保护方法

古籍数字化已经成为古籍研究的共识与方向,借助数字技术和科技手段可为中华优秀传统文化的弘扬与创新插上翅膀。然而,正如科技发展不可阻挡,越来越多的新技术也将不断丰富古籍数字化的技术领域。众所周知,数字化有狭义和广义之分,狭义的数字化代表了一个阶段,而广义的数字化将融入古籍保护和利用的全流程。因此,智慧性保护中的所有阶段都可以看成是古籍数字化的研究范畴,它没有固定的边界,并且会随着技术发展不断延伸。从古籍智慧性保护体系框架中可知,智慧性保护和再生性保护既相互关联,也各有侧重。保护与发展犹如 DNA 双螺旋结构中的两条主链,互相之间应不断产生共鸣。要实现发展中保护,保护中发展,就不能因循守旧,不能一味停留于现有的保护体系,需要结合实际,在现有的保护体系基础上不断创新保护模式。尤其是随着数字技术的发展,一切资源皆可数字化,一切资源皆可数字化连接。如何将更多维度的古籍资源关联起来,产生聚合甚至是化合作用,是古籍保护追求的目标和发展的方向。古籍智慧性保护并不是脱离古籍现有的原生性和再生性保护,而是起补充作用,为古籍保护提供新的思路 and 理念。

4.2 古籍智慧性保护贯彻了文化遗产保护发展理念

开放、融合、转化是文化遗产薪火相传、生生不息的鲜明标志。与其他文化遗产不同的是,古籍的价值不仅体现在古籍的外在形式,更与古籍的内容、与古籍所处的历史环境融为一体,不同时代、不同研究背景的学者对于古籍内容的解读和领悟也会产生变化,这些将赋予古籍保护整体性、动态性特征。“保护与发展并重,传承与创新并举”是中国特色古籍保护之路,智慧数据的显著特征就体现在知识产生、加工和积累的动态过程之中,这与动态保护的理念不谋而合。因此,需要在保护之初,制定好古籍保护工作的发展规划,充分考虑古籍内容的传递、知识的融通、历史的传承问题;建立多方合作的保护机制,秉承动态保护的理念,在古籍保护事业方面有所建树之后,为其他同类型文化遗产保护积累经验。古籍是中华优秀传统文化的重要载体和独特形式,如何对古籍进行保护、保存、利用和传承发展是一项重要的时代课题,在与国际文化遗产保护理念和保护方法接轨的同时,更要立足实情,形成符合自身文化特性和文化发展客观规律的保护发展理念。

4.3 古籍智慧性保护推动了古籍活化利用实现进程

保护是基础,利用是根本,一切古籍保护的目都是为了古籍能够合理利用、持久利用。浩如烟海的古籍资源,不应只是陈列于博物馆、图书馆的展品,或是被束之高阁的“书单”,它们应走进大众、融入生活,与时代结合、与发展同频。古籍活化的方式同样也需要不断地发展和创新,电子版、数字化是古籍活化,普及知识、演绎内容是古籍活化,在线展览、沉浸交互也是古籍活化。虽然本文提出的智慧性保护中的所有场景尚未完全得到实现,但其中不少方法已成为当前古籍研究的首选,如知识组织、古籍句读、实体识别、关联数据、知识图谱、知识计算等都已古籍保护和活化利用中发挥重要作用,

逻辑校验、语义推理、知识推理、知识表示、智能合约等研究方法也已经引起了学界的关注。本文提出的智慧数据驱动的智慧性保护框架,明确了活化传承的核心要义,构建了以古籍活化传承为主线的保护体系。借助该体系可弥补当前古籍资源保护中存在的不足,如数字技术应用过程中产生的阶段性数据,长期以来缺乏足够的重视和有效的保护,随着古籍智慧性保护的提出,这一问题有望得到改善。收集、整合古籍数字化进程中所有的相关信息,完成古籍知识分类体系,完善古籍知识理论体系,构建古籍智慧数据集,将为古籍的活化利用提供更为多元、更具智慧的数据来源。

4.4 古籍智慧性保护迎合了未来科技发展大趋势

数字时代,新技术的不断出现给各行各业带来了一次次的震撼,颠覆着人文领域的思维方式和研究范式,也深刻影响着文化遗产及古籍保护事业。2021年发布的“十四五”规划,将人工智能、区块链、大数据等新兴数字技术纳入其中。2021年兴起的元宇宙浪潮则来的更为猛烈,2022年3月,来自全国50家博物馆、高校的60位馆长、学者联名发布了《关于博物馆积极参与建构元宇宙的倡议》,便是为了让文化遗产“活”起来的重要体现。元宇宙打破了人们对互动媒体的理解,元宇宙技术也将颠覆或重塑人类的感知能力,而平面的古籍也会以更加个性化和沉浸式的新颖体验方式进行传承与传播,古籍也将变得更为鲜活和生动。另外,Web 3.0被认为是通往元宇宙的必经之路,它以分布式、语义网、区块链、人工智能、物联网等为核心特征。而智慧数据的实现过程,以及智慧性保护所提倡的知识范畴,都与Web 3.0的特征相一致,同时也与元宇宙中的BIGANT(区块链技术、交互技术、电子游戏技术、人工智能、网络及运算技术、物联网)技术体系高度同源。因此,可以预见,智慧数据驱动下的智慧性保护将会在未来科技发展的浪潮中发挥重要作用,为古籍

保护和利用提供无限可能。

5 结论与展望

本文在以下两方面做出了重要的尝试和探索。第一,梳理了智慧数据的技术体系,提出了从原生数据到语义数据、关联数据、智能数据和智慧数据的逐级进阶路径,为现有的智慧化应用提供了清晰的实施方案,如数字人文中的知识库构建、档案开放利用中的数智档案建设等。同时,也应认识到,智慧数据并非是一个静止的数据态,而应是一个动态流变的过程。第二,在智慧数据的驱动下,提出了古籍资源的智慧性保护体系。该体系将作为现有保护体系,尤其是再生性保护体系的重要补充。原生性保护主要保护古籍文物本身;再生性保护聚焦古籍的影印出版、缩微复制以及数字化形态;传承性保护强调古籍的活态演变;而智慧性保护的对象恰恰是古籍自身内容以及对外关联融合的多源知识元,这些将共同构成智慧性保护的知识

谱系。

古籍保护已不仅局限于原体的保护,数字化记录以及衍生出来的语义数据、关联数据、智能数据和智慧数据也应成为保护的重要手段。没有人文内涵的科技是冰冷的,失去科技赋能的传统文化也难以广泛传播。随着信息时代的到来,多元的数字形态、多维的保护体系,在完善古籍数字化信息的同时,也为将来古籍的管理、复原和活化利用提供重要依据和参考。智慧性保护如果可以在实现古籍保护目标的道路上成为不可或缺的一环,将势必推动古籍保护跨出关键一步,也会为古籍保护和我国的文化遗产保护事业提供新的研究方向。

当然,新技术的不断出现也将给古籍保护带来新的研究视角,笔者会持续关注并不断完善智慧性保护体系的内涵和外延。此外,本文所提出的智慧性保护框架仍然需要更多的古籍保护实践来指导,需要有更多的相关理论体系来论证,以期增加学界和业界对古籍保护的深入认识。

致谢:本文系国家社会科学基金重大项目“古籍保护学科建设与理论体系研究”(编号:19ZDA344)和古籍工作立项重点课题“古籍数字化资源平台互联互通研究”的研究成果。

参考文献

- [1] 单霁翔. 全球视野下中国文化遗产保护新发展[J]. 当代中国与世界, 2022(1): 53-68. (SHAN J X. The new development of China's cultural heritage protection from a global perspective[J]. Contemporary China and World, 2022(1): 53-68.)
- [2] 周飞亚. 让中华古籍真正“活起来”[N]. 人民日报, 2022-04-30(08). (ZHOU F Y. Let ancient books truly “come alive” [N]. People's Daily, 2022-04-30(08).)
- [3] 冯骥才. 建立国家非遗保护的科学体系[EB/OL]. (2020-09-29) [2022-04-15]. https://www.mj.org.cn/xwll/kmsd/202009/t20200929_231933.htm. (FENG J C. Establish a scientific system for the protection of national intangible cultural heritage [EB/OL]. (2020-09-29) [2022-04-15]. https://www.mj.org.cn/xwll/kmsd/202009/t20200929_231933.htm.)
- [4] 王新才, 聂云霞. 中国文献遗产保护的现状、问题与展望[J]. 图书情报研究, 2013, 6(1): 5-11, 24. (WANG X C, NIE Y X. The current situation, problems and prospects of the preservation of documentary heritage in China[J]. Library and Information Studies, 2013, 6(1): 5-11, 24.)
- [5] 周耀林, 姬荣伟. 文献遗产精准保护: 研究缘起、基本思路与框架构建[J]. 图书馆论坛, 2020, 40(6): 117-127. (ZHOU Y L, JI R W. Research on the precision preservation of documentary heritage: its origin, basic concepts and framework[J]. Library Tribune, 2020, 40(6): 117-127.)

- [6] 杨茜茜. 文化战略视角下的文献遗产保护与活化策略[J]. 图书馆论坛, 2020, 40(8): 163-172. (YANG Q Q. Documentary heritage preservation and revitalization policy: a perspective of cultural strategy[J]. Library Tribune, 2020, 40(8): 163-172.)
- [7] 王玉珏, 施玥馨. 联合国教科文组织文献遗产保护政策体系研究[J]. 图书馆建设, 2022(2): 120-130. (WANG Y J, SHI Y X. Research on UNESCO's policy system of documentary heritage protection[J]. Library Development, 2022(2): 120-130.)
- [8] 程焕文, 曾文. 国际图联的文化遗产保护理念与保护策略研究[J]. 图书馆建设, 2019(1): 47-54. (CHENG H W, ZENG W. Research on IFLA's cultural heritage protection concept and strategy[J]. Library Development, 2019(1): 47-54.)
- [9] 王秋云. 我国古籍数字化的研究现状及发展趋势分析[J]. 图书馆学研究, 2021(24): 9-14. (WANG Q Y. An analysis of the research status and development trend of digitization of ancient books in China[J]. Research on Library Science, 2021(24): 9-14.)
- [10] 苏祺, 胡切奋, 诸雨辰, 等. 古籍数字化关键技术评述[J]. 数字人文研究, 2021, 1(3): 83-88. (SU Q, HU R F, ZHU Y C, et al. Key technologies for digitization of ancient Chinese books[J]. Digital Humanities Research, 2021, 1(3): 83-88.)
- [11] 陈力. 数字人文视域下的古籍数字化与古典知识库建设问题[J]. 中国图书馆学报, 2022, 48(2): 36-46. (CHEN L. Digitalization of ancient books and construction of classical knowledge repository from the perspective of digital humanities[J]. Journal of Library Science in China, 2022, 48(2): 36-46.)
- [12] 刘忠宝, 赵文娟. 古籍信息处理回顾与展望[J]. 大学图书馆学报, 2021, 39(6): 38-47. (LIU Z B, ZHAO W J. Review and prospect of research on ancient Chinese text information processing[J]. Journal of Academic Libraries, 2021, 39(6): 38-47.)
- [13] 雷珏莹, 侯西龙, 王晓光. 数智时代古籍数字化再造的逻辑与进路[J]. 数字人文研究, 2022, 2(2): 46-56. (LEI J Y, HOU X L, WANG X G. The logic and approach of digital reconstruction of ancient books in the data intelligence era[J]. Digital Humanities Research, 2022, 2(2): 46-56.)
- [14] 李宗俊, 范炜. 面向智慧数据的语义丰富化方法及应用评述[J]. 情报科学, 2021, 39(7): 186-192. (LI Z J, FAN W. Review on smart data oriented semantic enrichment method and applications[J]. Information Science, 2021, 39(7): 186-192.)
- [15] 王晓光, 梁梦丽, 侯西龙, 等. 文化遗产智能计算的肇始与趋势——欧洲时光机案例分析[J]. 中国图书馆学报, 2022, 48(1): 62-76. (WANG X G, LIANG M L, HOU X L, et al. Intelligent computing of cultural heritage: a case study of European Time Machine project[J]. Journal of Library Science in China, 2022, 48(1): 62-76.)
- [16] SHETH A. Transforming big data into smart data: deriving value via harnessing volume, variety and velocity using semantic techniques and technologies[C]//30th IEEE International Conference on Data Engineering (ICDE). Chicago, USA, 2014.
- [17] LAFRATE F. A journey from big data to smart data[C]//Proceedings of the Second International Conference on Digital Enterprise Design and Management DED&M 2014. Paris, France, 2014: 25-33.
- [18] 王晓光, 谭旭, 夏生平. 敦煌智慧数据研究与实践[J]. 数字人文, 2020(4): 11-23. (WANG X G, TAN X, XIA S P. Research and practice of Dunhuang smart data[J]. Digital Humanities, 2020(4): 11-23.)
- [19] QIAN Y, XING Z Y, SHI X H. From collection resources to intelligent data: construction of intelligent digital humanities platform for local historical documents of Shanghai Jiao Tong University[J]. Digital Scholarship in the

- Humanities, 2020, 36(2): 439-448.
- [20] 郝伟斌,袁慧慧,王君仪. 文化遗产档案智慧数据资源建设研究[J]. 档案管理, 2021(6): 80-81. (HAO W B, YUAN H H, WANG J Y. Research on the construction of intelligent data resources of cultural heritage archives [J]. Archives Management, 2021(6): 80-81.)
- [21] 曾蕾,王晓光,范炜. 图档博领域的智慧数据及其在数字人文研究中的角色[J]. 中国图书馆学报, 2018, 44(1): 17-34. (ZENG L, WANG X G, FAN W. Smart data from libraries, archives and museums and its role in the digital humanities researches [J]. Journal of Library Science in China, 2018, 44(1): 17-34.)
- [22] 张永娟. 基于智慧数据的开放语义动态出版模式研究[D]. 上海: 上海大学, 2022. (ZHANG Y J. Research on open semantic dynamic publishing model based on smart data [D]. Shanghai: Shanghai University, 2022.)
- [23] 张云中,刘嘉琳. 智慧数据研究综述:概念辨析、价值取向、关键技术与应用框架[J]. 图书情报工作, 2021, 65(10): 141-150. (ZHANG Y Z, LIU J L. Smart data research: concept discrimination, value orientation, key technologies and application framework [J]. Library and Information Service, 2021, 65(10): 141-150.)
- [24] 李潇潇. 首次写入政府工作报告,期待点燃古籍保护热情[N]. 新京报, 2022-03-08 (A03). (LI X X. It was written into the government work report for the first time, looking forward to igniting the enthusiasm for the protection of ancient books [N]. Beijing News, 2022-03-08 (A03).)
- [25] 包弼德,夏翠娟,王宏甦. 数字人文与中国研究的网络基础设施建设[J]. 图书馆杂志, 2018, 37(11): 18-25. (BOL P K, XIA C J, WANG H S. The digital humanities and a cyber infrastructure for China studies [J]. Library Journal, 2018, 37(11): 18-25.)
- [26] 夏翠娟. 面向人文研究的“数据基础设施”建设——试论图书馆学对数字人文的方法论贡献[J]. 中国图书馆学报, 2020, 46(3): 24-37. (XIA C J. The construction of “data infrastructure” for humanities research: the methodological contribution of library science to digital humanities [J]. Journal of Library Science in China, 2020, 46(3): 24-37.)
- [27] 陈涛,祝蕊,苏日娜,等. 数字人文语义发布基础设施框架设计与实现[J]. 图书馆论坛, 2022, 42(8): 129-137. (CHEN T, ZHU R, SU R N, et al. Digital humanistic semantic publishing infrastructure: framework design and implementation [J]. Library Tribune, 2022, 42(8): 129-137.)
- [28] 韩寒. 数字化为古籍研究带来怎样的“蝶变”[N]. 光明日报, 2022-04-13 (08). (HAN H. What kind of “butterfly metamorphosis” does digitalization bring to the study of ancient books? [N]. Guangming Daily, 2022-04-13 (08).)

陈 涛 中山大学信息管理学院副教授,中山大学国家文化遗产与文化发展研究院研究员。广东 广州 510006。

苏日娜 中山大学图书馆馆员。广东 广州 510275。

张永娟 中国科学院上海生命科学信息中心副研究馆员。上海 200031。

张 靖 中山大学信息管理学院教授。广东 广州 510006。

余厚强 中山大学信息管理学院副教授。广东 广州 510006。

(收稿日期:2022-09-09;修回日期:2022-11-11)