建设文化遗产中文语料库的挑战与对策

徐拥军，陈晓婷，闫静

摘 要：【目的】文化遗产中文语料库作为文化大模型的关键组成部分，对于落实国家文化数字化战略、夯实文化大模型新基建、推动文化数据要素的市场化建设意义重大。【方法】运用文献研究法和归纳法，从理论层面探究文化遗产中文语料库的定义与类别，识别语料库建设面临的挑战，并提出相应的建设对策。【结果】当前，该领域面临高质量语料存量短缺、语料质量良莠不齐、语料标注标准不统一、语料数据权属不清晰等诸多挑战。【结论】建议将语料库建设纳入国家文化遗产治理体系，构建专门的国家级文化遗产语料库，建立多维度、高精度的语料质量评估体系，形成语义驱动、协同演化的语料标注机制，制定语料的数据共享与版权管理机制，以强化文化遗产中文语料对文化大模型的支撑力。

关键词：中文语料库；文化遗产；人工智能大模型；高质量数据；遗产治理

引用本文格式 徐拥军，陈晓婷，闫静．建设文化遗产中文语料库的挑战与对策 [J]. 创意设计源，2025(4):1-8.

Challenges and Countermeasures in the Construction of Chinese Cultural Heritage Corpus

XU Yongjun, CHEN Xiaoting, YAN Jing

Abstract: [Purpose] The Chinese corpus of cultural heritage is an important part of the cultural large language model. It is of great value for implementing the national cultural digitalization strategy, consolidating the new infrastructure of the cultural large language model, and promoting the marketization of cultural data elements. [Method] This paper adopted literature research and induction methods to examine the definition and category of the Chinese cultural heritage corpus from a theoretical level, identified the challenges faced by corpus construction, and proposed construction strategies. [Result] At present, this field faces multiple challenges such as insufficient high-quality corpus stock, uneven corpus quality, inconsistent corpus annotation standards, and unclear ownership of corpus data. [Conclusion] It is recommended to incorporate corpus construction into the national cultural heritage governance system, build a special national cultural heritage corpus, establish a multi-dimensional, high-precision corpus quality assessment system, form a semantically driven, co-evolutionary corpus annotation mechanism, and formulate a corpus data sharing and copyright management mechanism to strengthen the support of the Chinese corpus of cultural heritage for the cultural large language model.

Key words: Chinese corpus; cultural heritage; AI large language model; high-quality data; heritage governance

[**基金项目**] 本文系2024年度国家社科基金文化遗产保护传承研究专项“文化遗产治理体系和治理能力现代化研究”（项目编号：24VWB007）阶段性研究成果。

引言

在全球人工智能技术竞争加剧的背景下，语料库作为大模型构建的“数据地基”，其战略意义愈发凸显。近年来，我国高度重视语料库建设。2024年，国家数据局等17部门联合印发《“数据要素×”三年行动计划（2024—2026年）》，提出“挖掘文化数据价值，贯通各类文化机构数据中心，关联形成中华文化数据库，鼓励依托市场化机制开发文化大模型。”[1]2025年，中共中央、国务院印发《教育强国建设规划纲要（2024—2035年）》，提出建设“新型国家语料库”[2]。教育部、国家语委、中央网信办也提出，到2027年初步建成国家关键语料库[3]。然而，在全球主流人工智能模型中，英语语料长期占据主导地位，导致技术输出带有系统性文化偏见[4]。许多模型在涉及中国的历史文化、价值观时，存在系统性认知错误和偏见。文化遗产中文语料库内含高密度的中国文化价值观、高质量的语料，是引导我国大模型“向善”的关键。因此，如何建好、管好、用好文化遗产中文语料库，成为发展我国人工智能和保障文化安全的“双重命题”。

当前，语料库研究已从语言研究工具范式转向人工智能训练资源范式。中文语料库的发展历程从最初面向语言学研究的言语材料积累，向支撑知识挖掘与发现的

深度标注型知识资源方向演进[5]2。范式转型之下，学界逐渐确立了语料库的设计、采集、清洗、标注、存储与更新等流程体系，推动语料库由结构化语言数据走向多模态、高语义密度的智能资源库。人工智能时代下，中文语料库建设被提升到了国家战略高度。张凌寒强调大模型训练对高质量中文数据的依赖，提出应加快建设覆盖多场景、多领域的“可计算语料”[6]。李兴腾等指出当前存在“数据瓶颈”问题，呼吁构建国家级语料库运营平台以统筹数据供给[7]。与此同时，学者们也关注到生成式人工智能带来的语料版权风险[8]、训练数据的可追溯性与合规性[9]等问题。以上对建设中文语料库的呼吁，以及风险预判的相关研究，多聚焦于通用语料，尚缺乏针对文化遗产中文语料库建设的系统研究。

因此，本文基于建设文化遗产中文语料库的价值，梳理并归纳文化遗产中文语料库的定义、范围及其特点，剖析当前建设现状及其存在的问题，进而提出语料库的建设对策，为文化大模型、语言学研究等提供高质量数据支撑，系统保存和传承中华民族丰富的文化遗产资源。

一、建设文化遗产中文语料库的价值

文化遗产是国家和民族群体记忆的物质本源，关系到国家与民族的生存与发展，是一个国家、一个民族的命脉之所在[10]。建设高质量的文化遗产中文语料库，是中华优秀传统文化智能化表达的必要手段，将成为人工智能时代中华文明“可计算、可生成、可传播”的底层支撑系统，在构建中国特色、中国风格、中国气派的文化大模型中发挥不可替代的基础性作用。

（一）国家战略层面：实施国家文化数字化战略的要求

当前，全球人工智能竞争已从技术较量迈入价值观博弈的新阶段。中国工程院院士高文曾在公开演讲中指出，全球通用的50亿大模型数据训练集中，中文语料占比仅为1.3%[11]。暂且不论中文语料“量”的短缺通过翻译手段尚有解决方案，但中国文化价值观类语料相当重要却极为短缺[7]。大模型的表达能力和预测性能依赖于对大量知识和价值观数据的学习，而这类数据受人类主观价值的影响。古籍文献、历史文献等文化遗产语料是具有中国价值观的高质量语料，有助于大模型理解中文用户的文化背景和价值取向，保持中国文化自主性，服务中文用户。建设文化遗产中文语料库，有助于打破当前人工智能模型中西方知识和价值体系占主导的局面，改变西方知识体系在人工智能语料中的垄断地位。将中华优秀传统文化嵌入生成式人工智能的认知逻辑，构建契合中国自主话语体系与文化表达方式的智能模型，能够引导人工智能发展从“技术中立”迈向“价值可控”的战略转型，推动中华优秀传统文化融入数智时代的表达体系，切实达成文化主权可控、话语权增强的目标。

（二）科技创新层面：夯实人工智能文化大模型的新基建

文化遗产中文语料库是文化大模型的关键“食粮”，缺粮则智不强，数据质量决定模型未来。随着大模型向多模态、多语境、多场景方向发展，文化遗产中文语料从符号存储转向智能建模，成为生成式人工智能提升“文化理解力”“历史记忆力”的关键支撑，有助于构建中国自主的文化行业垂直大模型。通过对文化遗产语料的深度标注与结构建模，可支撑文化遗产实体识别与标注、关系抽取、知识图谱构建、语篇理解与重构等任务，形成具有中国气质的文化大模型与智能生成系统，这将进一步提高中华文化遗产语料在全球数字空间和人工智能关键应用场景中的比例与价值引领能力，同时提升其在全球语言治理框架中的参与度和在世界文明交流互鉴中的贡献率。

（三）经济产业层面：推动文化数据要素的市场化建设

数据作为第五大新型生产要素，蕴藏着巨大价值。在发展新质生产力的背景下，文化遗产语料正从“静态符号存储”转变为“动态数据资产”，从“信息载体”升级为“生产要素”。人工智能公司对高质量训练语料的需求日益迫切。在美国，语料库已变成“金库”，催生出庞大的人工智能数据交易市场。例如，为训练生成式人工智能模型，多家科技公司向图像托管网站Photobucket购买130亿张照片和视频，Shutterstock则将数亿张图片、视频和音频资源，授权给Meta、谷歌、亚马逊和苹果等巨头[12]。构建文化遗产中文语料库，形成“技术突破—场景落地—生态繁荣”的良性循环，不仅能够服务科研和大模型训练，也将成为“文化+科技”深度融合的核心数据资产。在应用层面，文化传播已呈现出智能化特征，大数据在文化活态体验、场景适配、跨媒介渠道、智能匹配等方面显示出强大的生命力[13]。语料库可广泛应用于数字文旅、智慧博物馆、非遗数字化、AI导览、智能对话机器人、文创设计等多元场景，打造面向大众的智能文化交互系统，丰富文化消费新业态。在产业层面，文化遗产语料库作为极具价值的数据资产，将为文化数据要素市场和文化科技融合产业提供有力的数据支撑，推动“文化资源—数字内容—智能服务”产业链实现高质量发展。

二、文化遗产中文语料库的定义、分类、特点

（一）文化遗产中文语料库的定义

联合国教科文组织和国务院颁布的相关文件将“文化遗产”界定为历史文物、历史建筑、人类文化遗址等“物质

文化遗产”，以及传统表演艺术、民俗活动、传统手工艺技能等“非物质文化遗产”的总和[14-15]。

“语料库”的定义随着技术环境的变迁而发生变化。1982 年，Francis 最早对语料库进行了定义，他认为语料库是一个用于分析语言的文本集合，是某种语言、方言或语言某方面具有代表性的文本集合[16]。早期的语料库主要依赖人工收集和分类，多从书籍报刊、公开演讲等文本中抽取，为学术研究提供了基础的语言结构、语法规则及词汇使用频率等数据[17]。之后，Sinclair 提出语料库是一个自然出现的语言集合，用于反映某一语言的状态和变化[18]。由此可见，早期的语料库主要是用来研究语言规律、发展和变化的一个集合。进入 21 世纪，随着技术的发展，语料库不再局限于传统的自然语言文本，还包括图文、影音等多模态数据。目前，大模型的训练语料主要来源于互联网、书籍、日常对话等通用文本，学术论文、代码库、百科等专用文本，以及图像、视频或结构化数据等多模态数据[19]。综上所述，语料库是数字化的、有一定规模的、能被计算机程序处理的语料集合[4]2。

中文语料库的定义应结合具体语境加以界定。从传统语言学的视角出发，狭义的中文语料库通常仅包括中文文本资源，主要服务于语言建模、问答系统、机器翻译等传统自然语言处理任务。而在人工智能多模态学习的背景下，其内涵进一步扩展，涵盖图像、音频、视频、三维模型等多模态中文语料，支撑图文生成、视频问答及多模态预训练等复杂任务。本文立足于人工智能时代的研究需求，所讨论的“中文语料库”是指为语言学研究或自然语言处理等技术应用系统，采集、整理并标注的多模态中文语料集合。

综上，本文将文化遗产中文语料库界定为在文化遗产相关活动中形成，经过数字化与规模化处理，能够反映其核心内涵、文化特征及其语境信息的多模态中文语料集合。该类语料库不仅包含文本、图像、音频、视频等多种数据类型，还具备数字化、规模化特征，能够支持计算机程序深度分析、语义挖掘及知识发现，从而为文化遗产保护、传承与创新研究提供数据基础与方法支撑。

（二）文化遗产中文语料库的分类

文化遗产中文语料库以中华优秀传统文化为核心内容，是文化语料库的重要组成部分。它通过数字化采集、结构化处理和数据标注，将古籍文献、历史文献、非遗文本、地方文献和民族语言文字等资源整合起来，形成可用于语言学研究、自然语言处理、大模型训练、文化传播的汉语语料资源体系。不同类型的文化遗产语料在来源场景、文化属性、数据形态、功能用途上存在差异。本文结合文化遗产语料的多模态特征与实际应用需求，提出语料组织与利用的分类标准（见表 1）。该标准不仅涵盖非物质文化遗产语料，也将与物质文化遗产相关的文字描述（如碑刻铭文、展品元数据、展览解说词等）纳入相应类别。

（三）文化遗产中文语料库的特点

1. 语料内容高度异质性，语言形式多样性。文化遗产语料来源广泛，涵盖古籍文献、碑刻契约、民俗传说、口述史料、展陈解说等多种形式。语言类型跨越文言文、方言、口语、民族语言等多种变体。在语言文字的流变过程中，还存在大量异体字、繁简混排以及特殊用语的情况。例如，一部古籍通常包括正文、注释、评语、版本对照等多个层级，同时还存在断句缺失、行文无标点、竖排右书等古代排版方式。这导致语料之间缺乏统一规范，分词、词性标注、句法分析的难度较高。

2. 多模态的融合特征。文化遗产在其自然形态中，以文字、图像、音视频、实物、动作等多种媒介形式，共同承载着文化意义的内在属性，形成跨媒介、多层次的信息表达体系。在文本之外，图像、音频、视频乃至三维模型在语料中占据重要地位。例如，敦煌遗书数据库[20]不仅包含敦煌文献的基本信息、全文录文和相关研究文献目录，还提供高分辨率的数字图像。数字敦煌资源库[21]包含敦煌经典洞窟的简介、朝代等基本信息，以及高清数字图像和石窟全景漫游的视频。敦煌学信息资源库[22]包含 334 738 条文博类资源，资源类型涵盖期刊、硕博士论文、电子书、图书、多媒体资源等。传统以文本为主的语料库构建范式已难以适配这一多模态生态，需要引入跨模态对齐、图文协同标注等新技术手段。

3. 文化语义的深度嵌入与语用复杂性，显著提升语料处理难度。文化遗产语料不是冷冰冰的语言符号，而是活态的、赓续中华文化脉络的话语标识，不仅传递事实性信息，更承载哲学、伦

表 1　文化遗产中文语料库的分类

类别	核心内容	主要来源	主要数据形态	代表数据库/平台
文献文书类语料	古籍、碑刻、契约、简牍、甲骨等传统文献	国家图书馆、敦煌研究院等	文本、图像	四库全书数据库、敦煌遗书数据库、数字简牍、“殷契文渊”甲骨文大数据平台
民族与地方文化类语料	族谱、地方志、口述史、民俗志、民族语言材料等	地方档案馆、文化中心、图书馆、民间组织	文本、音频、视频	温州“瓯越记忆”、温州古祠堂数据库、浙江志数字方志馆
非物质文化类语料	民间艺术、节日、手工艺、民歌戏曲等的记录与整理	非遗调查项目、访谈、视频采集、口述整理	文本、图像、视频、音频	中国非物质文化遗产数字博物馆、崔永元口述历史研究中心、中国戏曲志资源数据库
文化展陈与传播类语料	博物馆解说词、展品元数据、数字展陈文本、知识服务内容	数字博物馆、智慧文博平台等	文本、元数据 、多模态	上海博物馆数字文物库、故宫博物院数字文物库
学术研究与数字人文成果类语料	文化研究论文、专题数据库、知识图谱、多语种对齐语料等	高校、学术文献数据库、出版社、科研项目	文本、知识图谱、结构标注	文物出版社数字矩阵、知网智慧文博、“吾与点”古籍智能处理系统、数字敦煌资源库

理、历史记忆等深层文化意涵，具有意识形态属性。文化遗产中文语料库的词语使用、语篇结构体现着中华文化特有的思维方式，不能简单照搬西方语料处理范式。尤其在高语境文化背景下，这一复杂性更加显著。根据 Hall 提出的高语境与低语境文化区分[23]：高语境文化强调语境暗示与文化共识，对自然语言处理提出了更高的要求。诸如“礼”“仁”“义”“天命”等概念性词汇具有强烈的语境依赖性与文化指涉性，难以通过词频统计、关键词提取等浅层技术进行有效建模。低语境文化倾向于显性地表达不同。我国属于典型的高语境文化体系，语言所传递的信息高度依赖交际双方的文化背景、社会情境与历史知识。

4. 语料资源分布高度碎片化，数据权属与分类标准体系不清。长期以来，分散在文博机构、图书馆、档案馆、高校、科研院所、出版社、数据库商和各种项目中的文化遗产资源，呈现来源多元、形态丰富、获取路径复杂的特征。大量的文化遗产资源尚未数字化，少数已实现数字化的，也存在分类标准多样化、数据粒度参差不齐、版权归属不清、合规风险高等问题。据《中国文化遗产数字化研究报告（2023—2024）》显示，58% 的文博机构已迈出数字化采集第一步，但数字资产形式仍以二维图像和文本资料为主，目前文化遗产数字化应用还处于探索阶段[24]28。

三、当前文化遗产中文语料库建设面临的挑战

近年来，我国多个机构在文化遗产语料库建设方面已取得了积极进展。例如，国家图书馆的中华古籍资源库、文化和旅游部的在线博物馆平台、国家语言资源监测与研究中心的语言数据库等，为文化遗产中文语料库建设提供了良好基础。但整体来看，仍面临以下突出问题：

（一）高质量语料存量不足

尽管当前中文语料库总体规模可观，但真正可用于大模型训练的高质量语料仅占约 15%，其中超过七成来源于网络爬虫与论坛帖文，普遍存在语义冗余、噪声较大、质量参差等问题[25]。中华文化价值观类语料短缺已成为制约我国文化大模型发展的关键短板[26]，文化遗产领域的高质量中文语料更是稀缺。作为国家记忆的重要表达形式与大模型训练的关键资源，文化遗产语料正面临“高价值、低流通”的困境。长期以来，大量优质文化遗产语料被分散存储于不同文化机构、学术单位与地方项目中。数字化、数据化后的语料存量不足。主流大模型的训练数据缺乏稳定的、可控的文化遗产中文语料来源。据统计，国内几家主要古籍数字化机构全文数字化的古籍总量约为 168 878 种（部分为拆分的档案、敦煌文献），查重统计高达 91 516 种，去重后实际数字化的古籍仅约 7.7 万种（去除敦煌、档案等约 8 000 种 / 件）[28]4，而且仍以影像保存为主。

（二）语料质量参差不齐

文化遗产语料质量参差不齐，主要源于中文的复杂性和数字化处理的局限性。文言文、方言、诗词、碑刻及口述史等文本中，常见异体字、特殊排版、语句断裂等现象，对自然语言处理技术提出了更高要求。现有光学字符识别（Optical Character Recognition, OCR）系统对复杂文献的识别率低、错误率高，智能处理难度大，成为制约文化遗产语料向 AI 可用数据转化的关键瓶颈。例如，北京爱如生数字化技术研究中心的代表数据库“中国基本古籍库”，其全文数字化的古籍总量超过 5 万种，但文字差错率较高；北京翰海博雅科技有限公司的“鼎秀古籍全文检索平台”“文心阁古籍数据库”规模大，但文本质量低，版本错误多，检索不完善[27]4。此外，相关机构调研发现，超半数文博机构数字化仍以二维图像和文本资料为主[24]28，缺乏结构化、语义化、标准化的深度转化机制，限制了 AI 技术在文化遗产语境下的应用拓展。当前数字化工作多以“保存”为目的，尚未形成可计算、可调用的文化遗产语料资源体系，进一步加剧了“文化数据低转化率”的现实困境。

（三）语料标注的标准不一

数据标注是指对未经处理的原始数据添加说明、解释、分类或编码的过程，以便数据可以被人工智能算法所理解和使用。2024 年 12 月，多部联合印发《关于促进数据标注产业高质量发展的实施意见》，凸显了新一代高水平数据标注对于推动数据资源汇聚、提升数据质量、激活数据要素价值，支撑人工智能技术演进和应用落地的重要作用[28]。《时代》周刊（TIME）曾报道，为了在训练 GPT 模型的同时降低 ChatGPT 的有害性，OpenAI 曾以较低成本雇佣大量肯尼亚劳工开展数据标注工作[29]。大量人力投入是通用语料标注的基本要求。若要提升大模型的表达能力和预测性能，高质量专业人员的数据标注工作必不可少。Deepseek 的高性能得益于北京大学中文系学生参与的数据标注工作[30]。招聘平台发布的大模型数据标注岗位明确要求求职者具备与人类历史、文化、科学等相关的知识背景，以便协助数据工程师共同构建和完善世界语言知识库。

此外，文化遗产语料标注工作的复杂性远高于通用文本和一般互联网语料，具有较高的专业性需求。以古籍标注平台“吾与点”为例，其标注工作不仅涵盖文本结构标注、实体识别与关系抽取等，更涉及古籍编排规范、异体字处理、训诂注释等高专业性的知识标注，充分体现了对文化遗产专家知识与跨学科协作的高度依赖。然而，文化遗产语料资源分散，缺乏统一的数据标准

与协作机制，导致资源重复建设和数据孤岛的现象尤为突出，整体建设效率低下，特别在标注体系建设方面存在明显不足。现有标注标准体系尚不完善，虽然文化和旅游部于 2023 年发布了《非物质文化遗产数字化保护 数字资源采集和著录》系列行业标准，用于指导采集与基础著录工作，但数据的实体定义、语义标注、关系抽取、多模态对齐等深层次标注方面，仍缺乏统一规范，形成标准体系割裂、多级粒度混杂、跨机构数据兼容性不足等突出问题，亟须建立覆盖全流程、多层次、跨模态、统一的文化遗产语料标注标准体系。

（四）语料数据权属不清

目前，许多中文语料主要经由网络爬虫采集，数据溯源困难、归属不明，易引发版权争议和数据合规风险，限制了其在商业化和跨境合作中的使用。AIGC 同样存在侵权隐患，例如，谷歌被指使用 YouTube 视频的文字记录训练 AI 模型，涉嫌侵犯创作者版权 [31]。在我国，尽管中文语料资源相对丰富，但受《个人信息保护法》《数据安全法》等法律法规要求，工程师通常选择开源数据，但开源中文语料库也存在知识产权和隐私安全等潜在问题。2024 年，智源研究院在其发布的中文互联网语料库 CCI 2.0 使用协议声明中表示：审核语料库并不能保证完全不侵犯第三方知识产权，实际操作中仍可能存在侵权风险 [32]。

目前文化遗产语料库尚未形成统一的国家级数据资源统筹体系，缺乏统一的开放共享平台与分级可控使用机制，不同机构普遍存在“各建各的库”，语料难互通、难调用，跨机构调用和复用成本高昂。同时，部分语料来源不清晰、元数据不完善，缺乏完备的合规审核与授权管理，进一步限制了其在产业端的广泛使用与规模化商用落地。此外，部分文化遗产文本因历史归属复杂、著作权模糊或缺乏使用授权，使得语料无法进入商用或科研领域，影响其广泛推广和 AI 训练应用。阿里研究院在对比中美大模型后指出，我国高质量中文语料和科研数据的开放程度较低，企业用于模型训练的数据来源不明、权属不清，开源后存在合规隐患，致使多数企业倾向于“自采自用”，形成“孤岛化”建设现象 [25]28。

四、文化遗产中文语料库建设对策

（一）纳入国家文化遗产治理体系

1. 建立跨部门、跨系统的国家文化遗产语料库建设工作领导小组和协调机制。建议由中央宣传部、文化和旅游部、教育部或国家语委等关键部门牵头，统筹协调文物局、国家博物馆、国家图书馆、中央档案馆等成员单位，进行顶层设计、统筹协议。

2. 将语料库建设纳入国家文化遗产相关专项规划中。2022 年中共中央办公厅、国务院办公厅印发《关于推进实施国家文化数字化战略的意见》，明确提出，到“十四五”时期末，基本建成文化数字化基础设施和服务平台；到 2035 年，建成物理分布、逻辑关联、快速链接、高效搜索、全面共享、重点集成的国家文化大数据体系 [33]。这一过程中，文化遗产中文语料库的建设无疑是关键支柱。建议“十五五”时期，在有关文化强国建设、文化和旅游发展、文化产业发展、文化遗产保护传承、国家文化数字化等方面的规划、战略中，明确设立与“文化遗产中文语料资源建设”相关专栏或专项任务，作为支撑文化数字化与智能化发展的基础工程。通过专项引导，推动文化遗产中文语料库资源从零散采集向系统整合升级。

3. 将语料库建设纳入正在起草和修订的涉及文物和文化遗产保护的相关法律。2025 年 5 月，国务院办公厅发布的《国务院 2025 年度立法工作计划》中，明确提出预备提请全国人大常委会审议《非物质文化遗产法》（修订草案）、《文化产业促进法》（草案）、《历史文化遗产保护法》（草案），预备制定《历史街区与古老建筑保护条例》《传统村落保护条例》，预备修订《文物保护法实施条例》《历史文化名城名镇名村保护条例》等 [34]。建议在上述法律法规的后续修订中补充相关条例，推进文物资源数字化采集与中文语料整理，建设国家文化遗产语料资源体系，服务文化研究、公共传播和人工智能应用。例如，在修订《文物保护法实施条例》时，可在第二章“不可移动文物”的相关条款中补充：“鼓励将文物相关的文字、图像、语音等语言类材料系统收集，并转化为中文语料资源，纳入国家统一语料平台，形成支撑文化遗产数字化利用的中文语料库。”在第四章“馆藏文物”的相关条款中补充：“鼓励文物收藏单位对馆藏文物中蕴含的中文语料（如古籍、碑刻、口述资料等）开展语料数字化整理和标注，推动形成标准化的文化遗产中文语料数据库。”此外，针对文化遗产语料使用中日益凸显的数据确权、开放共享、治理标准等制度空白问题，建议在《文物保护法》中明确国家层面保障文化语料资源长期建设与规范使用的制度基础，明确制度性义务和权利边界。在《文物保护法实施条例》的“记录档案”“馆藏文物”等章节中，扩展“语料资源登记与利用”的操作性规范，补充包括语料采集的文字、图像、音视频等形式要求，语料编目与电子化管理流程，以及在语料登记、报备、借用、复制等环节中对权属标注、使用权限、平台接入等事项的具体规定。

4. 将语料库纳入文化遗产“系统性保护和统一监管”体系之中。当前，我国正从文化遗产保护向文化遗产系统治理转型，文化遗产中文语料作为“文化数据”的核心单元，应体现“系统性保

护和统一监管”的治理理念。在系统性保护层面，应坚持“整体性保护”理念，将语料资源与文物本体、历史环境、社会记忆等要素统筹纳入保护范围。同时，应推动语料资源与历史街区、非遗项目、学术成果等多源数据的融合关联，强化语料在文化遗产整体价值表达与认知建构中的基础作用。在统一监管方面，应明确文化遗产中文语料的法律地位与数据属性，制定采集标准、质量控制机制。并将语料资源视为文化遗产数据资产的重要组成部分，将其纳入文物监管和数据资产管理体系，实现从“保护物”向“保护数据”的治理延展。

（二）构建专门的国家级文化遗产语料库

在当前“国家文化数字化战略”与“文化大数据体系”建设的战略背景下，建议成立国家文化遗产语料库建设工作领导小组，统筹协调全国文化资源优势单位，推动构建专门面向文化遗产的国家级中文语料库。以多模态、多领域、多用途为导向，整合古籍文献、甲骨金文、简牍、碑刻、小篆等高价值文化遗产语言资源。

路径方面，借鉴已建立的 30 余项关键领域语料库的经验[35]，设立专门面向文化遗产中文语料建设的基金与专项资金，重点支持中文语料的系统采集、数字化整理、多模态标注与共享平台建设，鼓励跨学科协作与技术创新，推动构建覆盖范围广泛、结构清晰、服务多元的国家级文化遗产中文语料库体系。

管理机制方面，采用“中央统筹—地方执行—高校牵头—多方共建”的模式，由具备相关专业能力的高校与科研机构牵头建设，联合全国重点文博单位、非遗传承机构与数字人文实验室，形成职责明确、权责对等的协同网络。为避免推诿扯皮，应明确各参与方权责边界与工作节点，建立定期协调会议、项目进展通报与问题反馈机制，并设置第三方评估或考核指标。借鉴广东省在推动粤语语料库、古文字数据库等建设过程中形成的高校—研究机构—地方博物馆合作机制，探索区域试点与国家平台并轨推进的“双轮驱动”战略，鼓励跨学科、跨部门的紧密协作，增强语料库建设的系统性。

数据库内容构建方面，以建设文化遗产行业可信数据空间为导向，实现语料资源的规模化流通和共享，促进文化遗产数据要素的合规高效使用。突破以往以文本为核心的线性结构，迈向结构化、语义标注、多模态链接的知识型语料。通过构建包括图像扫描、释文标注、音视频转写、语义注释、知识图谱关联等多层级数据组织形式，实现语言资源从“数据孤岛”向“语义互联”跃迁。

（三）建立多维度、高精度语料质量评估体系

面对文化遗产语料在语言形式、表达方式、版本差异与媒介形态上的高度异质性，亟须建立一套科学、系统、可操作的语料质量评估体系，以解决“文化数据低转化率、低可用性”的根本问题。质量控制机制缺失、版本混杂、释文粗疏、标注不规范、结构不清晰等问题，难以满足科研建模、AI 训练、文化传播的多层次需求。为此，建议国家文化遗产语料库建设工作领导小组组织开展区域性试点和典型示范项目，以点带面推进国家文化遗产语料质量评估标准体系的建设与落地实践。

在内容维度，针对古籍文献，应评估其版本来源的权威性、文本的校勘质量以及注释的详尽程度；对于碑刻、甲骨、简牍等图像类语料，应评估其图像清晰度、字符辨识度以及释文准确性；对于非遗类语料，则应关注音视频资料的语言转写质量、文化术语准确性及语义的一致性。

在技术维度，评估语料的结构规范性、元数据完整性、标注统一性、主流文化大模型适配性。结合专家知识、大模型以实现语料的人机协作标注。例如，针对古籍文献中一词多义、通假字、避讳字等复杂问题，可构建领域专属的多义词本体库与上下文语境库。该库应基于历代辞书、注疏、版本对勘数据建立，结合上下文感知语言模型实现动态判别与标注。对于图像与释文对齐，采用“半自动标注—人工审核—模型迭代”模式，通过深度学习技术实现图文对齐。例如，运用图像分割、OCR，辅以手工标注，实现碑刻拓片、简牍、甲骨等图像型语料的像素对齐。

在流程维度，强调语料从采集、加工、审校到更新的全流程记录与版本控制机制，确保语料可溯源、可再现、可验证。具体机制设计上，建议引入专家同行评审机制，组建由语言学、历史学、文献学、古文字学等多学科专家组成的专业委员会，对重点语料进行逐级审核与分级认证。同时，可引入“文化语料质量星级制”或“语料可信度指数”等指标体系，以增强语料库试点成果的推广价值与跨机构通用性。

（四）形成语义驱动、协同演化的语料标注机制

文化遗产语料的高效利用，取决于其结构化与语义化加工的深度与精度。目前该领域普遍存在标注标准不一、标注粒度不清、标注对象与用途脱节等问题，直接制约了语料在跨平台、跨模态、跨任务中的流通与复用。为此，亟须构建一套面向多源异构文化遗产语料的语义驱动、可协同演化的标注机制。

为此，建议国家文化遗产语料库建设工作领导小组，在兼顾语料研究与人工智能训练双重需求的基础上，组织制定分类型、分层级的标注规范。围绕不同类型语料的语言属性与知识结构，明确标注范围、层级与方法，统一数据结构与字段格式，以满足学术研究的细粒度分析，同时支持 AI 模型的结构化学

习。例如，对古籍文献类语料，标注规范应覆盖基本文本内容、知识实体与语义关系，包括词性、句法依存、篇章结构等字段，以及人名、地名、官职、典籍、时间节点等文化实体字段，同时明确“典籍—人物—事件”的语义关系，构建符合知识图谱标准的实体与关系体系。对碑刻、简牍、甲骨等图像型语料，需实现图像与释文的对齐标注，记录图像坐标、文字位置索引、字体样式等。结合领域知识图谱与大语言模型，采用知识图谱定义文化遗产相关本体与关系类型，实现实体识别与关系抽取的人机协同标注，并通过专家审核，持续形成高质量的标注标准。

同时，鼓励有条件的高校开办“文化遗产智能计算”“数字人文”“文化遗产数字化”等专业（或专业方向），开设文化遗产语料标注相关课程，系统培养具备计算语言学、语言信息处理、数字文献学等知识的复合型人才。

（五）制定语料的数据共享与版权管理机制

在当前国际知识产权政策逐步向“促进创新”与“支持人工智能基础设施建设”倾斜的背景下，一些学者认为，大模型应用版权类语料进行训练，是为了掌握客观规律并培养模型基础能力，而非复制传播原作品、替代原有市场，可考虑将其“转换性”地合理使用或纳入法定许可范畴[25]30。如欧盟《单一数字市场版权指令》为符合条件的“文本和数据挖掘”设置了豁免例外[36]。日本政府在《著作权法》修改的同时，公开了自身对于版权法领域模型训练行为的态度——不会对 AIGC 模型训练中使用的内容加以版权保护。这些实践表明，训练用途与传播用途在版权法上应区别对待。我国应当积极探索文化遗产语料“转换性合理使用”的路径。

因此，建议国家文化遗产语料库建设工作领导小组推动制定适用于文化遗产语料大模型使用场景的专门版权与数据共享治理机制，核心内容如下：

一是明确要求凡接受国家财政资助的文化遗产项目所形成的中文语料数据，应坚持“开放为常态、不开放为例外”的原则，实现有序开放与共享。根据国务院办公厅 2018 年印发的《科学数据管理办法》[37] 和国家自然科学基金委员会自 2019 年起实施的科研数据开放政策[38]，政府预算资金支持下形成的数据资源，应向社会和相关部门开放共享。文化遗产中文语料作为国家重要的科研基础数据，理应纳入这一政策框架。建议组织编制文化遗产中文语料资源目录，建立统一规范的元数据标准，并将相关目录和数据及时接入国家数据共享交换平台，推动语料资源的系统汇交、管理、利用。

二是建立“文化语料可信共享联盟”，推动文博机构、高校实验室、全国文化大数据交易平台[39]、数据商、大模型企业等多元主体协同聚力，加入大模型语料联盟，推动可信共享、可控授权、合规溯源的语料服务生态建设，打破“自采自用、数据孤岛”现状。

三是按照数据价值贡献者的贡献大小获得相应报酬分配。2023 年 2 月，中共中央、国务院印发《数字中国建设整体布局规划》，提出开展数据资产计价研究，建立数据要素按价值贡献参与分配机制[40]。在完善初次分配、再次分配、第三次分配协调配套的制度体系基础上，应建立兼顾公共利益与权利保护的数据共享与版权管理机制，制定符合文化遗产中文语料特点的收益分配规则，保障数据要素收益在国家、权利主体、社会公众之间合理流动，共享文化遗产语料资源与开放利用成果。

文化遗产中文语料不仅是语言工程，更是文明工程，是推动中华优秀传统文化数字化表达、智能化传播、全球化影响的基础设施。当前正是推动文化遗产数据“从文献向语料、从语料向资产”转化的关键窗口期。国家和地方有关文化、文物、语言、教育、数据、网络主管部门，以及文博单位、高校、科研机构、出版机构等应积极协同参与构建权威、合规、高质量的文化遗产中文语料库，以推动中华文化在人工智能时代的创新表达与全球传播。

参考文献

[1] 国家数据局，中央网信办，科技部，等．十七部门关于印发《“数据要素 X”三年行动计划 (2024—2026 年)》的通 知 [EB/OL]. (2024−01−05)[2025−03−26]. https://www.cac.gov.cn/2024−01/05/c_1706119078060945.htm.

[2] 中共中央，国务院．中共中央 国务院印发《教育强国建设规划纲要(2024—2035 年)》[EB/OL]. (2025−01−19)[2025−03−26]. https://www.gov.cn/zhengce/202501/content_6999913.htm.

[3] 教育部，国家语委，中央网信办．教育部 国家语委 中央网信办关于加强数字中文建设 推进语言文字信息化发展的意见 [N/OL]. (2025−01−08)[2025−04−16]. https://www.gov.cn/zhengce/zhengceku/202503/content_7016543.htm.

[4] 钱明辉，杨建梁．高对齐数据集：人工智能新时代的文明守护 [EB/OL]. (2025−02−17)[2025−03−22]. https://m.jiemian.com/article/12356838.html.

[5] 黄水清，王东波．国内语料库研究综述 [J]. 信息资源管理学报，2021,11(3):4−17;87.

[6] 张凌寒．加快建设人工智能大模型中文训练数据语料库 [J]. 人民论坛・学术前沿，2024(13):57−71.

[7] 李兴腾，冯锋，黄鹂强．突破人工智能大模型的“数据瓶颈”：构建国家级语料库运营平台的思考 [J]. 中国科学院院刊，2025,40(3):522−529.

[8] 高雅文，来小鹏．生成式人工智能语料版权问题研究 [J]. 出版广角，2024(5):27−34.

[9] 邝苗苗，安小米，雷鸣，等．人工智能训

练数据真实性：概念体系构建及合规要求分析 [J]. 情报理论与实践 ,2025,48(7):65-73.
[10] 杨东 , 杨晋毅 , 杨茹萍 . 文化遗产保护的理论基础与三阶段论 [J]. 创意设计源 ,2019(1):29-32.
[11] 龚茜 , 何屹 , 房琳琳 . 大模型发展提速 中文语料够“吃”吗 [N]. 科技日报 ,2024-06-27(005).
[12] 第一财经 . 生成式 AI 热潮掀起“淘数据热”，背后风险有多大？ [EB/OL]. (2024-04-09)[2025-03-26]. https://news.qq.com/rain/a/20240409A05Q6R00.
[13] 宗立成 , 王娜娜 , 王雨曦 , 等 . 基于 AIGC 的传统文化创新转化路径研究 [J]. 创意设计源 ,2024(3):22-27.
[14] 联合国教科文组织 . 保护世界文化和自然遗产公约 [EB/OL]. (1972-11-23)[2025-08-25]. https://www.un.org/zh/documents/treaty/whc.
[15] 国务院公报 . 国务院关于加强文化遗产保护的通知 [EB/OL]. (2005-12-22)[2025-08-25]. http://www. gov.cn/gongbao/content/2006/content_185117.htm.
[16]FRANCIS W N. Problems of assembling and computerizing large corpora[C]//Computer Corpora in English Language Research. Bergen: Norwegian Computing Centre for the Humanities,1982:7-24.
[17] PACE-SIGGE M. Spreading activation, lexical priming and the semantic web: early psycholinguistic theories, corpus linguistics and AI applications[M]. Berlin:Springer, 2018:29-82.
[18]SINCLAIR J. Corpus, concordance, collocation[M]. Oxford: Oxford University Press, 1991:171.
[19] ZHAO W X, ZHOU K, LI J, et al. A Survey of Large Language Models[EB/OL]. arXiv preprint, arXiv:2303.18223, 2023. [2025-08-28]. https://arxiv.org/abs/2303.18223
[20] 敦煌研究院 . 敦煌遗书数据库 [DB/OL]. (2022-08-19)[2025-08-25]. https://dhyssjk.dha.ac.cn/.
[21] 敦煌研究院 . 数字敦煌资源库 [DB/OL]. (2016-05-01) [2025-08-25]. https://www.e-dunhuang.com/.
[22] 敦煌研究院 . 敦煌学信息资源库 [DB/OL]. (2024-08-27)[2025-08-25]. http://dh.dha.ac.cn/.
[23] HALL E T. Beyond culture[M]. New York: Knopf Doubleday Publishing Group, 1976:91-107.
[24] 腾讯研究院 .《中国文化遗产数字化研究报告》重磅发布 | “探元计划”2022 收官 [R/OL]. (2023-02-22)[2025-03-26]. https://news.qq.com/rain/a/20230222A076AP00.
[25] 安筱鹏 , 袁媛 , 宋志刚 . 大模型训练数据白皮书 [R/OL].（2024-10-22）[2025-09-11]. https://baijiahao.baidu.com/s?id=1813575825480921761&wfr=spider&for=pc.
[26] 李栋 . 建设新时代中国特色中文人工智能语料库的思考 [J]. 文献与数据学报 , 2024,6(3):27-33.
[27] 侯君明 , 陈媛媛 , 周佳益 . 中国古籍数据库资源建设的现状与展望 [J]. 天津师范大学学报（社会科学版）,2024(6):149-156.
[28] 魏亮 . 专家解读之二 | 繁荣数据标注产业，赋能人工智能高质量发展 [EB/OL]. (2025-01-15)[2025-03-26]. https://www.nda.gov.cn/sjj/zwgk/zjjd/0115/20250115110228353420038_pc.html.
[29] PERRIGO B. Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic[N/OL]. Time Magazine, 2023-01-18[2025-08-25]. https://time.com/6247678/openai-chatgpt-kenya-workers/.
[30] 阑夕 . AI 需要更多的热搜 [EB/OL]. [2025-05-31]. https://www.sohu.com/a/862951935_250147.
[31] METZ C, KANG C, FRENKEL S, et al. How tech giants cut corners to harvest data for AI[N/OL]. The New York Times, 2024-04-06(6)[2025-08-25]. https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.
[32] 北京智源人工智能研究院 . 中文互联网语料库 (Chinese CorpusInternet）使用协议 [EB/OL]. (2023-11-29)[2025-03-26]. https://data.baai.ac.cn/resources/agreement/cci_usage_aggrement.pdf.
[33] 中共中央办公厅 , 国务院办公厅 . 关于推进实施国家文化数字化战略的意见 [N/OL]. 2022-05-22[2025-05-12]. https://www.gov.cn/zhengce/2022-05/22/content_5691759.htm.
[34] 国务院办公厅 . 国务院办公厅关于印发《国务院 2025 年度立法工作计划》的通知 [N/OL]. 2025-05-14[2025-05-14].https://www.gov.cn/zhengce/content/202505/content_7023697.htm?utm_source=chatgpt.com.
[35] 丛芳瑶 . 教育部：已支持建设 30 余项关键领域语料库 [EB/OL]. (2025-04-01) [2025-03-26]. https://edu.gmw.cn/2025-04/01/content_37941506.htm.
[36] MAJUMDAR S. Directive (EU) 2019/790 of the European Parliament and of the Council: overhaul of European Union's copyright rules:a study[J]. Library Hi Tech News, 2020,37(9):11-13.
[37] 国务院办公厅 . 国务院办公厅关于印发科学数据管理办法的通知 [EB/OL]. (2018-03-17)[2025-05-12]. https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2018/201804/t20180404_139023.html.
[38] 张晓林 . 实施公共资助科研项目研究数据开放共享的政策建议 [J]. 中国科学基金 ,2019,33(1):79-87.
[39] 深圳文化产权交易所 . 全国文化大数据交易平台 [DB/OL]. (2022-08-31)[2025-08-25]. https://cn-cde.cn/Home.html?t=1744905296000.
[40] 徐心 . 推动数据要素收益全民共享 [N]. 光明日报 ,2024-12-09(09).

徐拥军，陈晓婷，闫静
中国人民大学