《文博语料库建设导则》
团体标准编制说明

一、 编制背景与必要性

当前，人工智能技术的飞速发展正深刻改变文化遗产保护、研究、展示与传播的方式，高质量、多模态的语料库已成为驱动“人工智能+文博”融合创新的核心基础战略资源。文博领域涵盖可移动文物、不可移动文物、博物馆业务及非物质文化遗产，其数据天然具有知识密集、多模态（文本、图像、三维模型、音视频）、时空关联性强、价值敏感等特点。构建符合领域特性的专业语料库，对于系统性保存文化遗产数字资源、深度挖掘其知识价值、赋能智慧文博体系建设、培育文化领域新质生产力具有至关重要的意义。

然而，我国文博语料库建设正处于快速发展与探索阶段，面临一系列挑战，亟需统一的标准规范进行引导：

1．方法论缺失与数据孤岛问题：各文博机构、研究单位和技术企业在语料库建设中，于数据模型、标注体系、技术流程上往往“各自为政”，导致大量重复建设，形成分散、异构的“数据孤岛”，资源难以整合与共享，制约了国家级文化遗产知识网络的建设。

2．多模态与知识治理挑战：文博数据涉及从藏品档案、考古报告、学术文献到高清图像、三维点云、修复视频等多种形态。如何系统性地采集、清洗、关联和组织这些多模态数据，并深入挖掘其内在的语义关系（如器物演变、历史事

件、人物谱系），缺乏体系化的技术路径指导。

3．智能化应用需求迫切：面向大语言模型、多模态AI模型的训练与微调，以及智能导览、虚拟修复、内容生成等具体应用场景，需要高质量、结构化、语义清晰的标注语料作为“燃料”。现有语料在规范性、深度和质量上参差不齐，难以满足AI技术对高质量数据的要求。

4．合规与可持续发展关切：文博语料涉及复杂的版权（如文物图像、研究成果）、个人隐私（如口述史）、文化敏感性（如涉及特定民族、宗教的内容）以及数据安全问题。如何合法合规地采集、使用语料，并建立可持续的运营与治理机制，是建设过程中必须解决的现实问题。

为此，积极响应《“十四五”文物保护和科技创新规划》、《“数据要素×”三年行动计划（2024—2026年）》等国家文化数字化战略，以及上海市打造人工智能高地、推动文旅元宇宙发展的地方政策，编制《文博语料库建设导则》团体标准，旨在为文博领域提供一套统一、科学、可操作的全流程建设规范。本标准的制定，将有效引导和规范文博语料资源的建设、治理与应用，打破信息壁垒，促进数据互联互通与价值释放，为人工智能深度赋能文化遗产事业提供坚实的标准化支撑。

二、 编制原则与依据

本标准编制遵循以下核心原则，并充分依据和参考了国内外相关标准、最佳实践及研究成果：

1．系统性原则：标准内容覆盖语料库建设的全生命周

期，从顶层设计（语料定义、核心知识模型）、生产过程（采集、清洗、标注、测试）一直到应用部署、安全治理与生态发展，形成完整的逻辑闭环。

2．领域适配性原则：紧密结合文博领域的知识特性，标准中专章定义了文博语料的多模态构成与核心实体关系，并设计了多层次、多粒度的标注体系，以精准刻画文物材质、工艺、年代、文化内涵等专业知识。

3．国际化与互操作性优先原则：积极倡导并融入国际文化遗产信息管理的主流标准框架。标准明确建议以 CIDOC CRM（文化遗产概念参考模型，对应国标 GB/T 37965-2019）作为核心语义模型的基础，并参考 LIDO（轻量级信息描述对象） 等元数据交换标准。这为确保不同机构构建的语料库未来能够实现语义层面的互联，融入全球文化遗产关联数据网络奠定了技术基础。

4．合规性与伦理性并重原则：设立独立章节专门规范数据安全、伦理与治理，强调版权审查与授权、数据分级与许可策略（如鼓励采用知识共享 CC 协议）、个人隐私脱敏、文化敏感性尊重等要求，确保语料库建设在合法合规的轨道上运行。

5．实践指导性原则：标准内容力求具体、可操作。在数据清洗、标注流程、质量测试等环节，提出了明确的流程步骤、方法建议和技术要求，可直接指导项目实施。

主要编制依据包括：

- 国际标准与最佳实践：ISO 21127:2014 (CIDOC CRM)，

LIDO Schema，欧洲数字图书馆 Europeana 的数据聚合实践，盖蒂研究所的 AAT（艺术与建筑叙词表）、ULAN 等权威受控词表。

• 国家标准与行业标准：GB/T 37965-2019《信息与文献 文化遗产信息交换的参考本体》、GB/T 42755-2023《人工智能 面向机器学习的数据标注规程》、GB/T 43697—2024《数据安全技术 数据分类分级规则》、文物行业相关数据采集与病害标准等。

• 基础性团体标准：T/SAIAS 015—2024《语料库建设导则》。本标准作为文博领域的专项导则，在通用语料库建设框架下，进行了深度的领域化扩展和细化。

• 前沿研究与项目实践：参考了故宫博物院构建“中国古代可移动文物概念参考模型（CRM-ACA）”的领先实践（如庄颖、叶祎珮的相关研究），以及国内外关于文物知识图谱构建、大模型在文博场景评测、高质量语料应用场景（如智能导览、AI 辅助修复、沉浸式体验）等方面的研究成果与案例总结。

三、 标准主要内容及技术要点说明

本标准共分为 11 章，结构上遵循“总则-过程-保障”的逻辑，核心内容说明如下：

第 5 章 文博语料：这是标准的基石章节。

• 5.1 多模态数据构成：明确了文博语料应由文本、图像、音视频、三维与交互式数据及元数据共同构成，并列举了每一类的具体内容（如文物描述文本、考古报告、高清文

物图像、修复过程视频、三维模型等），强调了数据的全面性与立体性。

• 5.3 核心实体与关系：提出了基于实体-关系（E-R）模型或知识图谱思想组织数据。定义了物质文化遗产对象、行动者、时间、地点、事件、概念六大核心实体类型，并示例了其间的核心语义关系（如“创作于”、“出土于”、“位于”）。这为将非结构化数据转化为结构化知识图谱提供了顶层设计框架，旨在从根本上解决数据关联性问题。

第 6-8 章 语料生产过程（采集、清洗、标注）：这是标准的核心技术流程部分。

• 第 6 章 采集：规定了数据来源应系统性覆盖机构内部权威资源、公开学术数据库、互联网开放资源（合规前提下）及数字化生产转化。提出了采集原则：系统性、质量优先、合法合规、可追溯、动态性。

• 第 7 章 清洗：规定了从“生语料”到“熟语料”的标准化清洗流程，包括格式统一、内容去噪、去重（文件级、段落级）、错误纠正（技术性、知识性）、文本规范化、数据脱敏及数据整合关联。强调了流程自动化与多级校验（自动化脚本+人工抽样）的质量控制机制。

• 第 8 章 标注：这是提升语料机器可读性与价值的关键。标准主张构建一个多层次标注体系：

• 基础语言学标注层：包括分词、词性标注、命名实体识别等通用 NLP 基础。

• 领域知识标注层：深度扩展，包括文博核心实体识别

与分类（文物、材质、工艺、朝代等）、关系与事件抽取、篇章功能标注。这部分内容紧密对接了知识库中“大模型在文博场景评测指标”对年代、材质、工艺、文化符号的准确性要求。

• 多模态关联标注：实现图文对齐、音视频-文本对齐、文本-3D 模型关联，是支持跨模态检索与理解的基础。

同时，对标注流程管理、规范制定、工具选型及质量控制（标注者间一致性评估、专家仲裁）提出了明确要求。

第 9 章 文博语料测试：为确保语料库交付质量，规定了通用测试与学科特色测试。

• 通用测试：包括输入输出规范性、文博概念准确性、多模态关联性、模型输入压力、语料分布覆盖性以及价值对齐检测。

• 学科特色测试：重点包括基于语料构建的文博领域知识图谱验证（结构合理性、关系准确性）、文博知识动态更新验证以及行业应用有效性检测。

• 人工检测：对检测流程和人员资质（初级、高级、专家）进行了分层规定，确保最终语料经过严格的专业审核。

第 10 章 文博语料使用：明确了语料库的应用价值方向，引导建设服务于应用。

• 学术研究与文化遗产保护：智慧科研辅助、文物保护专业辅助、行业数据基础设施。

• 公共服务与文化体验：智能导览与知识问答、在地文化创新体验（如 CityWalk）、沉浸式线下体验（VR 大空间等）。

• 数字内容高质量生成：多模态模型训练、影视游戏内容考证设计、XR 内容开发。此章节内容与知识库中“文博语料项目潜在应用场景”等文档的描述高度契合，体现了标准源于实践、指导实践的思路。

第 11 章 数据安全、伦理与治理：这是语料库可持续运行的保障。

• 版权与数据许可：要求事前审查，制定分层许可策略，鼓励采用知识共享（CC）等开放协议，并明确语料库作为汇编作品的知识产权。

• 隐私与伦理保护：严格执行个人信息脱敏，尊重涉及特定社群文化遗产的文化敏感性与权利。

• 数据安全与存储：要求全生命周期安全防护、数据备份与灾难恢复。

• 可持续治理与生态发展：建立持续维护更新机制、探索可持续的运营模式、鼓励社区参与与生态构建。

四、 与相关标准的协调关系

1．与 T/SAIAS 015—2024《语料库建设导则》的关系：本标准是该通用导则在文博专业领域的延伸、细化和具体化。在整体结构（如涵盖采集、清洗、标注、测试、安全）上与之协调，但所有技术内容均针对文博数据的特性进行了重定义和深度扩展。

2．与 GB/T 37965-2019 (CIDOC CRM) 的关系：本标准积极采用并倡导映射至该国际/国家标准。标准中“核心实体与关系”的设计思想与其一脉相承，并在“标准化映射与

语义发布”条款中明确建议将标注成果转化为基于CIDOC CRM的知识模型，以实现最高级别的语义互操作。

3．与GB/T 42755-2023等数据标注国家标准的关系：本标准在标注流程、质量控制等通用要求上与之保持一致，并在文博领域知识标注层的具体标签体系设计上进行了专业补充。

4．与文物行业现有数据标准的关系：本标准与文物普查、文物数字化采集（如WW/T 0114、WW/T 0115）、元数据（如DB11/T 1219）等行业标准是互补关系。行业标准侧重于基础数据的生产规范，而本标准则着眼于对这些基础数据及更广泛资源进行人工智能赋能下的二次加工、深度组织与创新应用。

五、 主要技术难点与解决方案说明

在标准编制过程中，重点研究和解决了以下技术难点：

1．如何定义文博领域的核心知识结构？

• 解决方案：没有凭空创造，而是复用和适配国际国内成熟的本体框架。以CIDOC CRM作为顶层逻辑，结合国内文博专家知识(如参考故宫CRM-ACA实践)和具体业务场景(如藏品管理、考古研究、展览教育)，提炼出“人、物、地、时、事、概念”六类实体及其关系。这既保证了语义的严谨性和未来的扩展性，又兼顾了国内的实际认知与需求。

2．如何处理多模态数据的深度关联？

• 解决方案：在标注体系设计中，设立独立的“多模态关联标注”层。不仅要求对各模态数据进行独立标注（如图

像目标检测、3D 模型部件标注），更关键的是强制要求建立跨模态的语义对齐关系，如将描述纹饰的文本片段关联到图像的具体区域多边形。这为构建真正意义上“可听、可视、可读、可互动”的统一文博知识体提供了技术路径。

3．如何平衡数据的开放共享与版权安全？

• 解决方案：在安全治理章节，提出了分层分级的管理思路。要求建设前进行版权状态审查，区分公共领域与受版权保护内容。鼓励对可开放内容采用知识共享（CC）协议明确授权。对于敏感或受限数据，通过数据脱敏（如对精确地理坐标、个人身份信息进行处理）、制定差异化的数据许可策略（开放访问、注册访问、受控访问）以及签订使用协议等方式，在促进数据价值流通的同时，筑牢安全与合规的底线。

4．如何确保标注质量并实现高效生产？

• 解决方案：提出了“人机协同、流程闭环”的标注质控体系。鼓励采用 AI 预标注技术提升效率，但必须辅以严格的人工多级审核（标注员自查、交叉校验、专家仲裁）和定期的标注者间一致性评估。同时，要求编制详尽的《文博语料标注指南》并动态迭代。测试环节的“模型输入压力检测”可将语料问题反馈至标注流程，形成持续优化的闭环。

六、 预期效益

本标准的制定和实施，预期将产生以下显著效益：

1．规范建设行为，避免资源浪费：为各类文博机构、技术企业提供统一的“施工蓝图”，减少因方法论混乱导致

的重复投资和低水平建设，推动行业资源向高质量、标准化方向集约。

2. 打破数据孤岛，激活要素价值：通过推广基于CIDOC CRM等标准的语义化组织方式，为未来不同语料库之间的互联互通奠定基础，促进文博数据要素在更大范围内的共享、融合与价值挖掘，助力构建国家文化遗产知识基础设施。

3. 赋能AI应用，培育创新生态：产出的高质量、结构化语料将直接服务于文博领域大模型的训练与微调，为智能导览、虚拟修复、内容生成、智慧研究等各类“AI+文博”应用提供可靠数据燃料，催生新业态、新模式。

4. 引导合规发展，保障可持续性：明确的安全、伦理与治理要求，有助于机构在建设初期规避法律与文化风险，探索合理的开放共享与可持续运营模式，确保语料库建设事业健康、长远发展。

综上所述，《文博语料库建设导则》团体标准是在国家文化数字化战略和人工智能发展浪潮下，应行业急需而编制的一份具有前瞻性、专业性和可操作性的重要技术规范。它融合了国际标准智慧与国内实践结晶，旨在系统性地解决当前文博语料库建设中的核心痛点，为我国文化遗产在数字时代的永久保存、深度研究与活化利用提供坚实的数据基石和方法论指引。