国家标准《信息技术 多场景、多字体藏文文字识别数据标注规范》（征求意见稿）编制说明

一、工作简况

（一）任务来源

根据《国家标准委关于下达2026年第六批推荐性国家标准计划及相关标准外文版计划的通知》（国标委发〔2026〕40号），由全国信息技术标准化技术委员会提出并归口的推荐性国家标准《信息技术 多场景、多字体藏文文字识别数据标注规范》制定项目（计划号为20263254-T-469）正式立项。

注：根据标准实际内容范围，计划名称《信息技术 多场景、多字体藏文文字识别数据标注规范》拟申请于报批阶段变更为《信息技术 多场景、多字体藏文文字识别数据标注》。

本标准由西藏大学主办，起草单位包括：西藏大学、电子科技大学、中国电子技术标准化研究院、青海师范大学、北京邮电大学。

本标准主要起草人：尼玛扎西、于永斌、仁青东主、黄姗姗、陈晓研、王向向、班玛宝、仁增多杰、头旦才让、洛桑嘎登、嘎玛扎西、拥措、群诺、范满平、陈安龙、道吉扎西、刘羽童、冯箫、沈嘉润、盛天颖、张子玥、次仁白玛、李臣波、郑智文、格桑加措、李蕾、李彝利、官却才让。

标准起草组人员组成和分工见表1。

表1 起草组人员组成和职责分工表

序号	起草人姓名	所在单位	工作内容
1	尼玛扎西	西藏大学	总体策划、组织协调、技术把关和统稿审核
2	于永斌	电子科技大学	技术方案牵头、核心条款及试验方案审核
3	仁青东主	西藏大学	藏文规则牵头、字体体系设计和技术审核
4	黄姗姗	中国电子技术标准化研究院	起草协调、资料调研、意见

			汇总和文本统稿
5	陈晓研	中国电子技术标准化研究院	标准化技术审查、规范性把关和材料组织
6	王向向	电子科技大学	识别技术路线、标注工具及试验方法编制
7	班玛宝	西藏大学	字体分类与字体类型标注规则编制
8	仁增多杰	西藏大学	数据采集清洗流程及质量控制要求编制
9	头旦才让	青海师范大学	文字内容标注及特殊字符规则编制
10	洛桑嘎登	西藏大学	场景分类、场景标注规则及示例整理
11	嘎玛扎西	西藏大学	文本框标注及异常图像处理规则编制
12	拥措	西藏大学	模糊残缺文本处理及转写示例校核
13	群诺	西藏大学	任务定义、需求变更及人员管理要求编制
14	范满平	电子科技大学	任务包创建和质量检查点要求编制
15	陈安龙	电子科技大学	账户权限、任务分发和数据安全要求编制
16	道吉扎西	西藏大学	藏文标注规则及字体内容一致性复核
17	刘羽童	电子科技大学	预标注试验、信息化检查及记录整理
18	冯箫	电子科技大学	标注工具测试和结构化格式示例校核

19	沈嘉润	电子科技大学	多场景样本整理及场景标签验证
20	盛天颖	电子科技大学	文本框、字体标注试验及问题统计
21	张子玥	电子科技大学	文字内容标注、字符编码及符号检查
22	次仁白玛	西藏大学	藏文语料审核和语言文字规范校对
23	李臣波	电子科技大学	相关标准调研、技术比对及文献整理
24	郑智文	电子科技大学	数据采集清洗验证及典型案例整理
25	格桑加措	西藏大学	典籍文献样本整理及标注规则验证
26	李蕾	北京邮电大学	自然场景与数字媒体样本整理验证
27	李彝利	电子科技大学	试验记录、征求意见及编制说明整理
28	官却才让	电子科技大学	标注试验支持、文本校对和材料归档

（二）制定背景

藏文多字体文本识别是扩展藏语语料数据、开展藏语自然语言处理研究的关键基础技术，直接影响智能文字处理系统的性能与泛化能力。作为跨区域使用的少数民族文字，藏文因其独特的字体多样性（包括乌金体、朱匝体、白徂体等传统字体，以及现代印刷体、手写体等），在国内外学术界和产业界存在巨大的技术需求，但目前仍缺乏系统性的多字体识别解决方案。

藏文作为我国重要的少数民族文字，其数字化处理能力直接影响民族地区信息化建设水平。据统计，西藏自治区政务系统中藏文文档占比超 40%，但现有 OCR 技术对乌金体、朱匝体等多字体的平均识别率不足 70%，亟需统一标准提

升识别效能。藏文古籍文献存量超过 10 万函，但数字化率不足 15%，标准缺失导致各机构采用自建标注体系，数据无法互通。在国际上，尼泊尔、不丹等"一带一路"沿线国家藏文使用人口约 200 万，缺少统一标准将阻碍跨境语言技术合作。制定《信息技术 多场景、多字体藏文文字识别数据标注》国家标准，能够为藏文文字识别、文字合成及相关应用提供统一的技术基础，满足藏族群众在教育、医疗、公共服务等领域的实际需求，推动藏族地区经济社会和信息产业的协同发展以及跨境语言技术合作研究。

我国政府高度重视少数民族语言文字的保护与发展，出台了一系列相关法律法规和政策规划。制定统一标准是对《国家语言文字事业"十四五"发展规划》中"加强少数民族语言文字信息化标准化建设"的落实，符合《新一代人工智能发展规划》关于"建立多语种智能处理标准体系"的部署，是对《关于推进实施国家文化数字化战略的意见》中文化遗产数字化保护任务的响应。

目前众多高校已经针对藏文文字识别进行了大量基础研究积累。西藏大学、西北民族大学等已建立包含 20 余万张标注样本的藏文多字体数据库，相关研究成果获 2023 年国家民委社会科学成果一等奖。基于深度学习的藏文文字识别技术已达实用水平。藏文文字识别标注已经具有较为完善的工具链，百度、科大讯飞等多家企业联合发起标准制定倡议，为主流 OCR 产品预留了标准接口。

本标准针对信息技术领域的藏文多字体文本识别需求，明确规定了图文数据的采集、清洗、标注规则及质量控制的流程与要求，涵盖印刷文档、手写文本、古籍文献、碑刻铭文等典型应用场景。本标准适用于卫藏、康巴、安多三大方言的文字数字化处理，包括但不限于文献数字化、实时翻译、内容审核、智能输入等自然语言处理任务。通过统一标注类型（如场景标注、字体标注、文本框标注、文本标注）、标注格式（如 JSON 标准化标记），本标准将有效解决当前图文语料库建设中存在的标注格式不兼容、标注标准不统一等问题，为全球研究者提供可互操作的语料基准。

（三）起草过程

2025年3月：起草核心人员编制完成标准草案预研稿。

2025年4月：全国信息技术标准化技术委员会字符集与编码分技术委员会（SAC/TC28/SC2）协助修改完善标准草案，准备全套标准立项材料。

2025年4月9日：参加全国信息技术标准化技术委员会组织召开的立项内审会，修改后同意立项。会后，起草核心人员按内审意见完善标准草案。

2025年7月11日：经全国信息技术标准化技术委员会下属藏文信息技术国家标准工作组审查，同意以此方案提出立项。

2025年10月：按照申请标准立项的要求，起草核心人员编制完善标准草案，由中国电子技术标准化研究院提出立项。

2026年6月27日：国家标准化管理委员会下达本标准制定计划。

2026年7月：西藏大学牵头成立本标准编制工作组，成员来自科研院所和高校等单位。起草组成员信息导入“标准智能研制系统”，分配系统账号。

2026年8月：编制形成工作组讨论稿（第1稿），经全国信息技术标准化技术委员会组织专家审查，同意开题，标准名称拟修改为《信息技术 多场景、多字体藏文文字识别数据标注》（于报批前提交更名申请）。会后，根据审查专家意见，就涉及数据安全的技术内容征求全国网络安全标准化技术委员会SAC/TC 260有关专家建议。

2026年9月：由起草组以线上会议形式组织召开标准编制讨论会，对标准草案进一步修改完善。起草组成员按照任务分工，通过“标准智能研制系统”对标准文本各章节协同编写，确保文本表述清晰、严谨，格式符合GB/T 1.1要求。

2026年9月底：形成征求意见稿初稿。

2026年10月14日：全国信标委组织召开内审会，审议通过标准征求意见稿初稿，并提出修改意见。

2026年10月至11月：由起草组以线上会议形式组织召开标准编制讨论会，根据内审意见，完善形成征求意见稿和征求意见稿编制说明。

二、国家标准编制原则、主要内容及其确定依据

（一）标准编制原则

本标准编制遵循规范性、科学性、适用性、可操作性和协调一致性原则，按照 GB/T 1.1—2020 的规定组织标准结构和技术表述，并与藏文字符编码、机器学习数据标注等相关国家标准相衔接。

一是坚持需求牵引、突出藏文特点。面向书籍、手写笔记、石碑铭文、数字

化媒体和自然场景等应用，充分考虑藏文字体类型多、字形差异大、文本图像质量不一等特点，设置场景、文本框、字体类型和文字内容四类标注任务，使技术内容能够服务于藏文图像文字识别和目标检测等实际需要。

二是坚持全流程设计、注重可操作和可追溯。技术要求覆盖待标注数据采集与确定、任务定义、标注规则、需求变更、环境管理、任务创建与分发等环节，明确数据需求方、标注管理方和标注方的职责，并通过任务说明、质量检查点、审核反馈和记录归档形成闭环。

三是坚持技术中立、开放兼容。标准以功能要求、标注规则和数据字段为核心，不限定必须采用某一具体标注产品；草案中的工具界面仅作为实现示例。文字内容采用 GB/T 13000—2025 规定的字符编码体系，文本框位置、字体标签和转写内容采用结构化字段表达，以提高数据在不同工具、平台和应用之间的交换与复用能力。

四是坚持安全可控、持续改进。对数据采集、传输、存储、标注和分发过程提出权限管理、脱敏或加密、日志记录和监控等要求；对需求变更设置评审、预标注、验收和重新确认程序，并通过小规模预标注、信息化检查、人工复核和争议仲裁不断校正标注规则。

（二）标准主要内容

本标准规定了信息技术领域多场景、多字体藏文文字识别数据标注的要求。本标准适用于藏文图像文字识别、文本检测等处理。

本标准的主要技术内容涵盖多场景、多字体藏文文字识别数据标注的全流程，包括流程设计、前期准备和任务执行，具体如下：

（1）图文数据采集标注流程

图文数据采集标注流程由五个核心环节构成：图文数据采集、图像数据清洗、数据标注、标注数据检验与审核、质量评估。首先，采集原始图文资料，并进行数据去重、图像噪声处理和图像旋转裁剪。随后，由经过专业培训的标注人员按照统一的标注规范完成标注（标注规范依据本文件规定的标注规则和数据需求方的具体要求制定）。所有标注数据均应进行文本框检验和文本检验，并接受审核。审核不合格的，应将问题反馈给标注人员，重新标注并检验，直至审核合格。审核合格的数据应由数据质量评估团队进行质量评估，评估通过后方可交付给数据

需求方

（2）标注任务前期准备要求

- 待标注数据的采集及确定：明确了待标注数据的来源、采集方法和质量要求，确保数据输入的可靠性和一致性。

- 图文数据标注任务定义：以SunOCR标注工具为例，进行场景标注（标注图像中的场景，如书籍、广告牌、手写笔记、石碑等）、文本框标注（标注文本框中心点坐标，以及长宽像素值）、字体类型标注（标注每个文本框中文本所使用的字体类型，如乌金体、簇仁体、簇通体、朱匝体、柏簇体、丘伊体）、文字内容标注（标注每个文本框中对应的文字内容）。

- 标注规则：详细描述了场景标注、文本框标注、字体类型标注、文字内容标注的具体要求。

- 标注需求变更流程：规定了当数据需求方提出需求变更时应遵循的流程，确保需求管理的有效性。

- 标注环境要求：规定了标注工具的选择标准和标注场地的环境要求，确保工作条件符合技术需求。

（3）标注任务执行过程的要求

- 标注任务创建要求：包括任务包创建和任务账户创建两个环节。任务包创建需对数据按类型、内容或难度合理分组，设置质量检查点和明确标注说明；任务账户创建需根据角色配置账户权限，制定账户使用规则，确保数据安全和标注规范性。

- 标注任务分发要求：需根据任务要求选择合适的分发方式（如主动领取或系统自动分配），制定优化策略平衡质量与效率，并采取安全保障措施包括建立安全管理制度、实施日志记录与实时监控，确保分发过程的安全合规。

除上述主要内容外，本标准另设资料性附录“标注人员管理要求”，明确了标注人员的选拔范围、培训要求及能力档案管理，确保人员素质达标。

（三）标准主要内容的确定依据

本标准主要内容是在藏文文字识别数据建设的实际需求、藏文文字和字体特征、现有标注工具能力以及相关标准化成果的基础上确定的。起草组以提高多来

源藏文图像数据的一致性、准确性、可交换性和可追溯性为目标，按照“对象可描述、过程可执行、结果可检查”的思路设置各项技术要求。

范围和术语的确定主要依据藏文 OCR 训练、测试及目标检测等任务的共同需求。标准将对象限定为信息技术领域多场景、多字体藏文文字识别数据，并通过“图文数据标注任务”“图文数据需求方”“图文数据标注管理方”“图文数据标注方”等术语明确业务边界和参与角色，为后续流程要求提供统一语义基础。

采集标注流程和前期准备要求主要依据数据生产实践确定。针对原始图像可能存在重复、噪声、方向不正、模糊、残缺以及敏感信息等问题，草案规定采集、整理分类、清洗、脱敏或加密、任务定义、人员培训和环境准备等要求；场景类别覆盖书籍、手写笔记、石碑铭文、数字化媒体和自然场景，以兼顾典籍数字化、政务文档、教育资料和公共标识等典型应用。

四类标注任务及其规则依据文字识别模型所需的核心监督信息确定。场景标注用于表征图像来源和应用环境；文本框标注采用左上角坐标及宽、高描述文字区域；字体类型标注覆盖乌金体、簇仁体、簇通体、朱匝体、柏簇体、丘伊体以及“其它”类别；文字内容标注采用 GB/T 13000—2025 编码，并对模糊文本、残缺文本、无法辨认字符、特殊符号和多文种混合等情况规定统一处理方式。上述内容能够满足检测、分类和识别任务对位置、类别和转写文本的基本数据需求。

任务执行和质量控制要求依据规模化数据标注的管理需要确定。通过任务包分组、质量检查点、角色账户和权限配置、主动领取或系统自动分发、日志记录与实时监控等措施控制过程风险，保证技术要求能够被验证和追溯。

在标准协调方面，草案规范性引用 GB/T 13000—2025 以统一藏文字符编码，并参考 GB/T 42755—2023 关于机器学习数据标注的一般流程，以及 GB/T 20001.5—2017 等标准化文件编写要求。同时结合 SunOCR 等现有工具的界面和结构化输出示例验证字段可实现性，但标准条款保持对具体工具和实现方式的中立。

三、试验、验证情况分析

（一）试验情况

本文件的试验验证以技术路线的可实施性、标注规则的可理解性和全流程的

可贯通性为重点。起草组结合草案列举的书籍、手写笔记、石碑铭文、数字化媒体和自然场景等代表性图像，以及乌金体、簇仁体、簇通体、朱匝体、柏簇体、丘伊体等字体类别，对数据准备、任务定义、标注、检查、审核和归档环节进行了示例性流程验证。

在标注实现方面，利用草案所示 SunOCR 工具对场景标签、文本框、字体类型和文字内容四类信息进行关联标注，检查了文本框左上角坐标及宽、高的记录方式、字体标签选择、藏文转写录入，以及 bbox、label、transcription 等结构化字段的导出形式；同时检查标注工具对 GB/T 13000—2025 藏文字符显示和编辑的支持情况。该过程用于验证各类标注对象和数据字段能够在现有工具环境中落地。

在规则适用性方面，围绕图像旋转矫正、最小外接文本框、单行文本框、字符间距过大时拆框、不同字体分框、重叠文字、模糊或残缺文本、无法辨认字符、特殊符号和多文种混合等典型情形进行逐项检查。通过正反示例和统一占位标记，使同类异常数据具有一致的处理路径，便于标注人员理解和审核人员复核。

从现有示例和流程检查情况看，草案规定的四类标注任务能够形成场景、位置、字体和文字内容相互关联的结构化结果，任务创建、分发、检查、审核和归档能够构成闭环，技术路线具备可实施性。现阶段验证主要用于支撑标准技术条款设计，尚未形成大规模、多单位条件下的统计性试验结论。后续将结合征求意见和实际应用，扩大不同场景、字体、图像质量和标注人员的验证范围，记录一致性、错误类型、返工情况和工具兼容性等结果，并据此进一步完善标注规则和试验方法。

（二）预期的经济效益

本标准适用于文化机构开展古籍数字化工作，支持政府单位实现政务文档智能化处理，助力企业开发 OCR 技术和翻译软件，同时为学术机构进行藏语自然语言处理研究提供技术规范。通过规范产品、服务和流程，标准将推动藏文文字识别技术在各领域的规范化应用和产业化发展，使藏文 OCR 产业规模进一步扩大，大幅降低企业二次开发成本。

（三）预期的社会效益

本标准实施后将推动民族地区电子政务、在线教育等公共服务水平提升，惠

及600余万藏文使用人口。同时，通过藏文文字识别技术可加速濒危古籍文献抢救性数字化，为珍贵文献进行标准化建档。

四、与国际、国外同类标准技术内容的对比情况

本标准技术内容为我国独创，无同类国际标准、国外标准。

当前国内外尚未形成针对藏文文字识别标注的专项标准体系。在国际标准层面，英语、汉语等主要语种已形成完善的文字识别标注规范，如《信息处理 机器可读字符编码（MICR和OCR）》（ISO 2033:1983）。在国家标准中，中文OCR标注规范也已形成，如《人工智能 面向机器学习的数据标注规程》（GB/T 42755-2023）。同时针对主要语种也具有大量的开源标注工具，如PPOCRLabel、Label Studio等。然而，针对藏文文字识别的标准化工作仍面临关键挑战：多字体标注体系缺失、多场景适应标准空白、结构化标注框架待完善。本标准通过建立统一的标注规范，将显著提升藏文OCR系统的跨场景识别率，推动形成以我国为主导的少数民族文字智能处理标准体系，填补该领域国际标准空白，为"数字丝绸之路"建设提供重要的标准支撑。

五、以国际标准为基础的起草情况

无同类的国际、国外标准，本标准未采用国际、国外标准。

六、与有关法律、行政法规及相关标准的关系

本标准未替代或废止现行相关标准，与现行相关法律、法规、规章和强制性国家标准协调一致。

本标准规范性引用了GB/T 13000—2025《信息技术 通用编码字符集(UCS)》,规定在藏文文字识别数据标注应采用GB/T 13000的字符编码体系，以保障藏文字符在跨系统、跨国界交换传输中的一致性；规范性引用了GB/T 45674—2025《网络安全技术 生成式人工智能数据标注安全规范》，以保障待标注数据的安全性。

本标准在起草过程中还参考了以下我国国家标准：

- GB/T 42755—2023 人工智能 面向机器学习的数据标注规程

七、重大分歧意见的处理经过和依据

无。

八、涉及专利的有关说明

本标准不涉及专利。SunOCR 标注工具是本标准起草单位西藏大学尼玛扎西院士团队用于多场景文字识别的软件，目前处于研发阶段。该工具在本标准草案中仅作为标注工具示例，用于说明多场景、多字体藏文文字识别数据的标注流程和结构化输出方式，不作为实施本标准的必要条件。SunOCR 标注工具目前不涉及专利和商标。

九、实施国家标准的要求以及措施建议

建议本标准作为推荐性国家标准发布，自发布之日起 6 个月实施。在标准实施中，建议及时在行业内及相关单位开展标准的宣贯，使相关单位能够学习并应用本标准，并结合实际应用要求贯彻实施标准。

十、其他应当说明的事项

无。

国家标准《信息技术 多场景、多字体藏文文字识别数据标注规范》编制工作组

2026-11-18