ICS 35.240.99
CCS L70

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIAS 032—2025

规划与自然资源语料库建设导则

Guidelines for the Construction of Corpus on Planning and Natural Resources

2025-07-24 发布　　2025-07-25 实施

上海市人工智能行业协会　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由上海市人工智能行业协会提出并归口。

本文件起草单位：上海市规划和自然资源局、上海库帕思科技有限公司、上海市人工智能行业协会、上海市数字城市规划研究中心、工业互联网创新中心（上海）有限公司、同济大学、上海商汤智能科技有限公司、浙江时空智子大数据有限公司、上海脉策数据科技有限公司、上海市城市规划设计研究院、上海市测绘院、上海市自然资源调查利用研究院、上海市城市建设档案馆、上海市建设用地和土地整理事务中心、上海市自然资源确权登记事务中心、上海市土地交易事务中心、上海市测绘产品质量监督检验站、上海城市规划展示馆、上海市地矿工程勘察（集团）有限公司、上海市规划和自然资源局执法总队、上海市土地储备中心、上海市政工程设计研究总院（集团）有限公司、上海市城市建设设计研究总院（集团）有限公司、同济大学建筑设计院（集团）有限公司、北京超图软件股份有限公司、上海市园林设计研究总院有限公司、华东建筑集团股份有限公司、清华大学建筑设计研究院上海分院、上海同济城市规划设计研究院有限公司、上海谈瀛数字科技中心、联通（上海）产业互联网有限公司、中国电信股份有限公司上海分公司、中城交科技（上海）有限公司、上海浦东发展（集团）有限公司、上海城建城市运营（集团）有限公司、联通数据智能有限公司、上海工业自动化仪表研究院有限公司、国创智造科技（上海）有限公司。

本文件主要起草人：张玉鑫、山栋明、孙珊、王训国、许健、徐明前、宋唯、庄澜、黄海清、奚文沁、施佳樑、张文佳、程蓉、徐玮、钟俊浩、董云皓、汪旻琦、肖扬、徐祎、郭羽、汤舸、杜展展、俞佳炜、饶雪、赵春昊、张琳、蒋钦、欧阳琳欣、李光雨、徐望悦、吴菁妍、徐政、冉江、夏凉、张一鸣、忻静、任晨阳、冯威丁、范华、刘映、王军、刘爽、罗梦婷、毛施施、李晨曦、詹武明、赵薇、陈四平、郑雅珊、徐子捷、周鑫栋、郭忠诚、张斌、任中佳、杨海涛、吕倩、赵颖、王峰、徐元玮、温南南、沈璐、刘婷婷、程晓楠、堵炜炜、虞祝豪、朱莉琴、杨文恺、郑更河、华静、宋佳琪、王强、李轶承、龚伟、张来勇、郭汉杰、魏飞、金常飞、常光照、周赛赛、刘通、孙颖、赵环宇、王汇文、金恩、滕丽、戴彬、潘登、赵兴华、陈锡铭、潘登、赵兴华、陈锡铭、郭爱华、沈彦、肖红练、贺仁龙、郑茂宽。

本文件首次制定。

首期执行单位：上海市规划和自然资源局、上海库帕思科技有限公司、上海市数字城市规划研究中心、工业互联网创新中心（上海）有限公司、同济大学、商汤集团有限公司、浙江时空智子大数据有限公司、上海脉策数据科技有限公司、上海市城市规划设计研究院、上海市测绘院、上海市自然资源调查利用研究院、上海市城市建设档案馆、上海市建设用地和土地整理事务中心、上海市自然资源确权登记事务中心、上海市土地交易事务中心、上海市测绘产品质量监督检验站、上海城市规划展示馆、上海市地矿工程勘察（集团）有限公司、上海市规划和自然资源局执法总队、上海市土地储备中心、上海市政工程设计研究总院（集团）有限公司、上海市城市建设设计研究总院（集团）有限公司、同济大学建筑设计院（集团）有限公司、北京超图软件股份有限公司、上海市园林设计研究总院有限公司、华东建筑集团股份有限公司、清华大学建筑设计研究院上海分院、上海同济城市规划设计研究院有限公司、上海谈瀛数字科技中心。

引　言

为适应“人工智能+”在城市治理以及规划与自然资源领域的应用要求，推进高质量、标准化语料资源的建设，制定本文件。本文件规定了规划与自然资源行业基础语料的组织、分类、处理与管理要求，明确语料库建设的基本原则、主要内容与实施路径，用于支撑量子城市等新型基础设施建设对高质量语料资源的应用需求。

规划与自然资源行业基础语料是指在行业业务活动中形成的、具有典型时空特征的文本、图像、音频、视频等数据集合，可用于支撑大模型训练、智能识图、语义检索、城市模拟与辅助决策等智能化应用场景。

本文件建设原则包括：

一是全面系统、统筹集成。语料库建设应覆盖规划设计、自然资源、测绘地理等核心知识领域及其交叉内容，依托结构化知识树体系，系统整合行业知识内容，保障语料在知识点层面的覆盖完整性与语义一致性。

二是质量保障、时空精准。语料由业务部门、科研机构、高校和专业语料服务单位联合采集，数据应来源明确、内容准确、表达规范。应结合业务流、知识流与空间流组织语料内容，形成可映射至二维图纸与三维模型的表达结构。

三是边建边用、持续迭代。语料建设可分阶段推进，优先覆盖应用频次高、需求紧迫的知识领域。在知识体系支撑下，采用“边建边用”的组织策略，动态更新语料内容，涵盖行业政策、技术方法与业务体系更新形成的新数据资源，构建支持多方协作与持续更新的语料资源体系。

四是合法合规、安全管理。语料采集、处理、存储与共享应符合国家法律法规和行业规范要求，落实数据脱敏、权限控制与信息安全机制，保障语料资源的合规使用和系统运行安全。

本文件所构建的基础语料资源可为规划和自然资源行业大模型体系建设提供语料支持。所提供的结构化文本、图像标注、遥感信息及空间知识可用于支撑如下能力建设：

——为行业通用大模型提供语义检索、语图理解与智能推演所需语料；

——为专业模型提供遥感分析、城市问题诊断、空间生成、指标评估、资产登记等任务训练素材；

——为政策解读、规划辅助、模拟评估、指挥调度等业务场景提供结构化数据支撑。

本文件的实施可支撑构建统一规范、来源可靠、语义一致的行业语料资源体系，为规划与自然资源领域的智能化发展奠定基础，为数据驱动的量子城市与城市治理体系提供基础保障。

规划与自然资源语料库建设导则

1 范围

本文件规定了规划与自然资源语料的语料库分类、数据、语料生产及其治理与安全等要求。

本文件适用于规划与自然资源领域语料库的建设、管理、应用与评价等工作。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

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

GB/T 45396—2025 数据安全技术 政务数据处理安全要求

GB/T 45574—2025 数据安全技术 敏感个人信息处理安全要求

GB/T 45577—2025 数据安全技术 数据安全风险评估方法

T/SAIAS 015—2024 语料库建设导则

3 术语和定义

3.1

数据资源 data resources

以电子化形式记录和保存的具备原始性、可机器读取、可供社会化再利用的数据集合。

[来源：T/SAIAS 015—2024，3.1]

3.2

语料 corpus

语言材料或语言应用的样本。

[来源：T/SAIAS 015—2024，3.3]

3.3

语料库 corpora

由依据一定抽样方法收集的自然出现的语料所构成的电子数据库。

注：是按照一定目的和方法进行选择并有序排列的数据汇集。

[来源：T/SAIAS 015—2024，3.4]

3.4

模态 modal

机器对现实世界信息的感知模式或信息通道，包括数据表征模式（例如文本、图像、语音、视频、生物和生理信息的数据表征）、数据采集机制（将每种传感设备采集到的数据视为一种模态），以及数据特征主体（如对特定主体的局部信息进行数据化表征）。

[来源：T/SAIAS 015—2024，3.5，有修改]

3.5

敏感信息 sensitive information

如果公开或者滥用会造成潜在危害的信息。

[来源：GB/T 4894-2009，4.7.3.2.4，有修改]

3.6

去标识化 de-identification

通过对个人信息的技术处理,使其在不借助额外信息的情况下,无法识别或者关联个人信息主体的过程。

注:去标识化建立在个体基础之上,保留了个体颗粒度,采用假名、加密、哈希函数等技术手段替代对个人信息的标识。
[来源：GB/T 35273-2020，3.15]

3.7

匿名化数据 anonymized data

去除直接涉及数据主体的个人或组织数据。

[来源：GB/T 4894-2009，4.7.3.2.3，有修改]

3.8

规划与自然资源语料库 corpus on planning and natural resources

国土空间规划、测绘地理信息、自然资源调查监测评价、自然资源统一确权登记等规划与自然资源领域相关文本、图片、音频、视频等语料库。

4 缩略语

以下缩略语适用于本文件。

AI：Adobe Illustrator 文档 (Adobe Illustrator Document)
BMP：位图图像格式 (Bitmap Image Format)
CDR：CorelDRAW 文档 (CorelDRAW Document)
CoT：思维链 Chain of Thought
DOC：文档文件格式 (Document File Format)
DOCX：文档扩展格式 (Office XML Document Format)
D0C：文档文件格式 (Document File Format)
ECW：增强压缩波格式 (Enhanced Compression Wavelet Format)
EPS：封装 PostScript 格式 (Encapsulated PostScript Format)
GIF：图形交换格式 (Graphics Interchange Format)
HTM/HTML：超文本标记语言 (HyperText Markup Language)
JPG/JPEG：联合图像专家组格式 (Joint Photographic Experts Group Format)
KHz：千赫兹 (Kilohertz)
MP4：动态影像专家压缩标准音频层面 4 (Moving Picture Experts Group Audio Layer IV)
OCR：光学字符识别 (Optical Character Recognition)
PDF：便携文档格式 (Portable Document Format)
PNG：便携式网络图形格式 (Portable Network Graphics Format)
RTF：富文本格式 (Rich Text Format)
SHP：ESRI Shapefile 格式 (ESRI Shapefile Format)
SFT：监督微调 Supervised Fine-Tuning
SVG：可缩放矢量图形 (Scalable Vector Graphics)
TIFF：标签图像文件格式 (Tagged Image File Format)
TXT：文本文件格式 (Text File Format)
WAV：波形音频文件格式 (Waveform Audio File Format)
WEBP：网络图像格式 (Web Picture Format)
XML：可扩展标记语言 (Extensible Markup Language)

5 语料库分类

5.1 业务分类

5.1.1 管理类语料库

管理类语料库采集主体为规划和自然资源管理局各业务处室、行业主管部门或协会，采集日常管理和运作过程中生成和积累的业务资源，包括但不限于政策法规、技术标准、业务规则、管理案例、审批成果等。

——政策法规聚焦相关领域的规范性文件，包括法律、行政法规、部门规章、地方性法规、地方政府规章、国务院和部委规范性文件、其他规范性文件等。

——技术标准包括国家标准、行业标准、地方标准、团体标准、企业标准等。

——业务规则包括规划资源管理部门牵头制作的办事指南、业务相关的重要概念释义和其他指导业务工作开展的非发文业务规则。

——管理案例侧重收集规划资源管理部门在管理工作中积累的、用于指导管理工作的典型案例。

——审批成果包括规划资源一张图管理系统中沉淀的各类审批流程数据和审批成果数据。

5.1.2 学术类语料库

学术类语料库采集主体为国内规划与自然资源领域各高校、研究机构、职业考试培训机构、出版社，采集权威且广泛应用的教材、职业培训与考试资料，以及企事业单位内部技术指导文件。包括但不限于学术著作、学科教材、培训用书及期刊论文等。

——学术著作主要涉及国土空间规划理论、城乡发展规律及相关交叉学科著作。

——学科教材包括规划、建筑、测绘等专业方向的教材资料。

——培训用书覆盖注册规划师、注册建筑师等职业考试材料。

——期刊论文内容覆盖政策创新与技术前沿，具有学术价值与工程实践导向。

5.1.3 实务类语料库

实务类语料库采集主体为规划和自然资源业务相关实务单位，采集日常运作过程中生成和积累的业务资源，包括但不限于包含实务案例、技术手册、书籍期刊等。

——实务案例分为科研课题、研究报告、获奖项目、方案征集和其他。科研课题包括立项的研究课题；研究报告包括行业研究报告、单位自研报告等，如普查报告、统计年鉴、皮书、智库成果等；专业奖项包括本市及外省市、全国城乡规划奖、国土空间规划奖和其他领域获奖项目等。

——技术手册聚焦用于实务培训的技术指南和操作手册。

——书籍期刊指实务部门自行生产的专业书籍。

5.2 用途分类

5.2.1 预训练语料库

预训练语料库是支撑规划与自然资源行业模型初始训练的大规模无标注数据集合，涵盖文本、图像、音频、视频等多模态信息，用于通用表征学习。

5.2.2 监督微调语料库

SFT语料库是针对规划与自然资源行业特定任务优化模型的定向数据集合，由标注的任务相关样本构成，用于引导模型掌握特定场景下的响应逻辑和行为规范。

5.2.3 思维链语料库

CoT语料库是用于训练和评估规划与自然资源模型在处理包含多种信息模态的复杂问题时的推理能力的数据的集合，包含问题、显式推理步骤与最终答案。

5.2.4 知识库语料库

知识库语料库是在大模型训练过程中应用的辅助性语料资源集合，外挂接入并用于承载领域内非结构化数据的向量数据库，支持在模型推理阶段或应用阶段进行精准查询和增强。

6 数据要求

6.1 数据属性要求

6.1.1 基本数据信息

规划与自然资源语料库的资源数据包括文本数据、图像数据、音频数据、视频数据，语料表达详见附录A。

6.1.2 数据规模要求

数据资源应具备充分的数据体量，需满足各业务场景下模型训练、算法验证及实际应用的最低数据量要求，具体数值应根据不同业务场景的实际需求确定。

6.1.3 数据多样性要求

数据资源应包含多元性的数据格式与种类。

6.1.4 数据密级要求

作为数据资源最小组成单元的数据文件，应不设置加密、编辑限制、访问控制等内容操作权限限制。

6.1.5 数据质量要求

数据资源应具有高质量和可靠性，不含水印标识及不适宜内容。

6.2 数据技术要求

6.2.1 数据提供过程要求

在数据提供前，应明确数据来源、采集方式及使用目的。

对原始数据进行处理或转换时，应保留原始数据副本，并记录处理流程和操作说明。

在将数据提交至语料资源库或共享平台之前，应对数据格式、结构、内容进行统一检查。

若涉及多源异构数据，应进行标准化预处理，包括但不限于格式转换、元数据补充、数据清洗等。

6.2.2 数据传输及存储要求

数据的提供、传输与存储应保障其安全性、完整性与准确性，防止数据在传输过程中丢失、损坏或被非法访问。传输通道应具备一定的带宽和稳定性，宽带要求根据数据的类型、规模及传输的实时性需求确定，稳定性要求传输通道的丢包率与延迟波动应不超过实际可接受的数值。

对于存储环节，应选择可靠的存储介质和系统架构，支持数据的长期保存与快速调用。数据存储格式应符合通用性行业要求，优先采用通用性强、兼容性高的格式，格式包括但不限于以下类别：

——文本数据优先采用XML、JSON、TXT格式。

——图像数据优先采用JPEG、PNG格式。

——音频数据优先采用MP3、WAV格式。

——视频数据优先采用MP4、AVI格式。

6.2.3 数据质量要求

所提供的语料数据应满足高质量标准，具体包括：准确性、完整性、一致性、可溯源性、代表性与可移植性。数据内容真实反映实际场景或业务情况，避免人为干预导致偏差。

数据应附带完整的元数据信息，包括但不限于数据来源、采集时间、采集方式、处理方法、坐标参考系统、精度指标等。

7 语料生产要求

7.1 语料生产路径

语料生产技术路线应执行“采-洗-标-测-用”的全流程管理，确保语料资源的实用性、准确性和可用性。

——采集环节应系统性收集规资领域的原始数据，避免语料缺失和偏差。

——清洗环节应采用自动化方式去除噪音、冗余和无效信息。

——标注环节应采用人工与自动化相结合的方式，提升语料准确率。

——测试环节应通过采样检测，验证语料的完整性和一致性。

——应用环节应建立反馈循环以迭代优化语料质量。

针对规资领域语料的多样性特点，预训练、SFT、CoT及知识库等不同用途语料的生产，其路径选择深度融入“采洗标测用”全流程管理要求。

——预训练语料生产，可无标注环节，测试环节应验证预训练模型的通用语义理解能力。
——SFT语料生产阶段，标注环节应结合人工标注与自动标注，测试环节应验证模型对目标任务的泛化能力。
——CoT语料生产阶段，标注环节应进行逻辑结构标注，测试环节应验证模型链式推理的准确性与合理性。
——知识库语料生产阶段，可无清洗、标注、测试环节，采集环节应采集非结构化与结构化数据

7.2 采集

7.2.1 采集数据类型

数据采集宜按规划与自然资源领域知识体系采集，知识体系见附录B。

7.2.2 数据命名与标签

统一数据格式和命名规则，统一文件命名规则为：统一标识码+文件名称。
对各类数据材料进行知识库标签管理，标签管理请参见附录C。

7.2.3 数据适用环境

按照GB/T43697—2024的数据分级规则的数据属性准备不同的语料加工环境。

7.3 清洗

文本数据的清洗，需检查完整性，填补或标记缺失值；校验准确性，纠正错误内容；统一格式、单位和命名规则，消除不一致性；检测并删除重复数据；处理异常值和噪声数据。

图片数据的清洗，使用图像处理算法去除噪声，统一转换为JPEG、PNG等格式，调整尺寸，删除重复图像。

在清洗过程中，推动数据标准化和规范化，确保数据的互操作性和可扩展性。

建立自动化清洗流程和工具，提高清洗效率和效果。

重视数据安全和隐私保护，采用数据脱敏，包括但不限于以下技术：

——地图与图形脱敏:专题图、地形图、规划图、遥感图像等地图类数据在公开或跨域使用前，需采用包括但不限于坐标扰动、加密投影、比例尺限制、分辨率裁切、图层筛选、地理遮蔽、图斑置换等在内的脱敏技术手段。
——文字与文档脱敏:文字与文档类数据需采用包括但不限于命名实体识别脱敏、正则表达式批量替换、上下文语义遮蔽等在内的数据脱敏技术手段。
——遥感数据脱敏:遥感数据在使用时应采用包括但不限于区域加密脱敏、遥感影像去特征处理等在内的数据脱敏技术手段。

7.4 标注

7.4.1 标注原则

标注体系应具有扩展性，能够根据新数据类型和应用需求不断更新。

在标注过程中，应注意保护用户隐私，对敏感信息进行脱敏处理。

应建立详细的标注指南，明确标注标准和流程，定期评估和更新标注规则。

7.4.2 标注路径

语料标注路径应包括认知过程维度与提问策略维度。

认知过程维度聚焦概念与要素认知、语义与空间理解、应用与分析评价等角度设计问题，对应规划与自然资源领域从基础知识记忆到迁移应用的认知层次提升。

提问策略维度基于封闭式、开放式、比较式等各类问题角度针对性提问，通过差异化、高密度提问强化对知识内容的多维理解。

7.4.3 自动化标注

自动化标注通过研发适用于规划与自然资源语料生产的自动化数据加工处理流程，宜符合GB/T 42755-2023的数据标注规程。

7.4.3.1 标注方法

自动化工具利用预训练模型和规则引擎，高效处理大规模、规则明确、重复性高的任务，基于基础文件进行解析，快速生成预标注结果。

自动清洗技术手册、业务流程文档，拆分“问题-思维链-回答”结构，利用大模型生成强推理语料。

7.4.3.2 标注内容

在文本数据标注方面，需确保标注精确性和一致性，使用统一的术语和标签，标注关键信息，如规划指标、资源类型、政策条款等。

对于图片数据，需标注关键信息，如地质构造、土地利用类型、建筑轮廓等，同时提取并标注图像中的文本内容。

对于音频数据，需将其转换为文字信息，并按照文本数据标注要求进行处理，确保规划与自然资源行业术语和语义的准确标注。

对于视频数据，需进行切片分段处理并提取关键视频帧，参照图片数据的标注规范进行关键信息识别与标注。

标签标注形式采用结构化标签体系，为各类数据赋予标准化标签。文本标注使用关键词标签与语义类别标签，图片采用区域框选标注结合属性标签。

7.4.3.3 标注审核

应明确标准化校验规则，覆盖格式、逻辑、语义；应自动识别异常并标记，联动人工复核，形成闭环校验流程。

7.4.4 人工标注

7.4.4.1 标注方法

针对重要知识领域、重点项目和重点业务类型进行文本对、图文对、思维链、图片标注等人工微调语料的生产。

文本对主要针对重要知识领域开展，包括概念、理解、应用、评价、分析等问答类型。

图文对主要针对绘制类图像和拍摄类图像开展，包括类型、要素、空间关系、政策关联、CoT推理、CoT评价等问答类型。

思维链主要针对重点项目、重点业务类型开展，挖掘项目编制背后的经验与逻辑内容，包括问题、思维过程和答案。

图片标注主要针对服务于特定场景的绘制类图像和拍摄类图像，应结合场景识别需求开展图面标注。

7.4.4.2 标注内容

文本对和图文对的内容包括问题和答案，采用人工撰写问题、自动生成答案、人工核验答案、循环修正的多轮闭环方式。问题围绕知识点提出，具有明确清晰、专业性，包含背景和核心诉求。答案自动生成后，应人工进行打分和修改。

思维链的内容包括问题、思维过程和答案，采用全人工撰写的方式。问题为强推理问题。思维过程步骤清晰，表达简洁。答案应按照思维过程逻辑开展，形成一一映射。

图片标注的内容包括类型、要素（图名、图例、指北针、符号、编码等）、空间关系等方面，采用人工框选和文字标注的方式。

完成人工标注后，应为每条语料内容添加知识树标签，明确语料对应的知识领域。

7.4.4.3 标注审核

规范标注审核流程，人工语料的审核可采用业务审核、行政审核两级审核机制，确保其兼顾规划和自然资源实务和管理领域的使用需求。

审核内容应重点聚焦提问错误、概念错误、逻辑混乱、答非所问、知识缺漏等问题。

7.5 存储

对于语料存储平台或语料资源库，应建立完善的数据安全机制，包括设置访问权限、源文件加密、自定义数据格式等方式。

应定期开展语料备份与恢复演练。

应考虑语料公开属性，建立公开级与受限级语料加工环境。参照GB/T43697—2024分级要求，设置公开与内部两类存储环境。

——公开语料存储：主要负责加工可以在公有云环境存储的数据。包括学术类（行业领域的学术著作、学科教材、培训用书）、实务类（法律法规、规范性文件、技术标准、局主动公开的规范性文件、局认可的国内外优秀案例）和管理类（可公开的规划与自然资源业务数据）。

——内部语料存储：主要负责加工在专网并实施权限管控的数据，具体包括管理类和实务类。其中，管理类涵盖项目基本通则、项目体系规范以及不适宜公开的规划资源业务运行数据等；实务类包含项目业务规则、项目操作法则、相关研究报告以及非主动公开的规范文件等。

7.6 测试

数据测试包括数据质量评估和应用质量评估两方面。

数据质量评估重点关注文本、图片、表格数据的准确性、一致性、完整性和及时性，通过量化指标如准确率、完整率等进行监测。

建立含多维度评测体系的标准测试集，设置基础量化指标与行业特性扩展指标。

7.7 应用

7.7.1 模型预训练

基于法律法规库、标准规范集、历史档案库等基础语料，通过海量行业文本的泛化学习，形成模型对国土空间规划、资源配置、生态保护等全领域知识的底层认知框架。

7.7.2 模型后训练

7.7.2.1 监督微调

基于SFT语料，针对用地审批登记、规划条件核发等标准化业务场景，专项优化模型对具体指令的输出合规性与格式精度。

7.7.2.2 推理训练

基于CoT语料，引导模型模拟专业推理逻辑，形成可解释的分析链条。

7.7.3 知识检索增强

规划与自然资源知识库以语料库为数据底座，通过知识抽取、结构化处理，形成涵盖法规标准、技术规范、案例库等内容的知识体系。

7.7.4 智能体应用

规划与自然资源智能体依托语料库知识储备，通过自然语言处理技术实现人机交互，可应用于业务咨询、方案预审、公众答疑等场景。通过语料资源服务的形式强化智能体应用，结合实际业务需求优化对话逻辑与应答策略，确保交互内容符合行业通用要求及业务场景普适性原则。

7.8 更新

应建立语料的周期性更新制度，通过技术手段与机制创新实现，定期采集新发布文件充实语料库，包括但不限于以下材料：

——政策法规、技术标准类文件宜定期追踪采集；

——学术类经典著作等材料宜年度更新；

——专业期刊类文件宜按月更新。

应通过优化成果规范、提交要求等，在各类材料编制中同步生产高质量人工文本对、图文对等内容，并随材料同步归集至语料库。

8 治理与安全要求

8.1 治理要求

8.1.1 管理体系

应构建整合性治理的管理体系,促进不同业务与专业领域简的沟通协作；应优化组织管理流程，实现对语料从采集源头到最终应用的全过程管控。

语料治理框架应紧密结合国土空间全域管控，由自然资源主管部门牵头，协调生态环境、住房和城乡建设、交通运输、农业农村、民政、文化旅游、商务、应急管理等多部门成立联合治理委员会。

8.1.2 全生命周期治理

应围绕“全生命周期”展开语料治理，明确覆盖采集、处理、应用各环节的规范性要求。

8.1.3 可持续运行

应定期开展语料质量评估，利用自然语言处理模型检测文本语料的逻辑矛盾，例如规划文件中相互冲突的指标，通过图像相似度算法排查重复或低质图片，动态优化语料库内容。

8.1.4 生态化协作

应搭建行业语料贡献激励平台,鼓励科研机构、规划设计院上传脱密脱敏后的研究报告与案例图片，贡献者可通过积分兑换数据调用权限；应与高校合作建立标注联盟，制定文本语义标注规范、图片分类标准，统一语料标注规则。

8.2 安全要求

8.2.1 接入安全

在语料接入阶段，如涉及传感器、终端设备或其他外部接口，应采取加密手段（硬件/软件加密）对语料进行保护。确保仅授权用户可获取相应数据。

对外提供的语料应进行脱敏处理，尤其是涉及个人隐私、商业秘密或国家安全的信息，应按照GB/T 45574—2025进行过滤或屏蔽，涉及政务语料安全，应符合GB/T45396—2025的要求。

8.2.2 传输安全

语料通过限制传输范围保障安全，需确保处于内网加工环境，不接入公共网络。

8.2.3 应用安全

所有语料应明确提供方与使用方的权属关系，必要时应签订使用协议，界定用途与使用范围。

应建立语料访问日志机制。

9 价值要求

语料产品数据应具有正确的价值导向，符合科技伦理和社会道德。

——语料采集环节的数据来源应按照GB/T45577—2025进行评估风险，避免虚假信息或价值观偏离。

——语料清洗环节应符合GB/T 45574—2025的要求进行清洗。

——语料标注环节应按照标注基准与行业道德等规则对齐。

——语料测试环节应进行价值一致性验证，拦截违反社会公平的决策逻辑。

——语料应用环节应构建道德迭代机制，要求动态修正语料中的伦理缺陷。

附 录 A
（资料性）
规划与自然资源语料表达示例

A.1 基本数据信息

规划与自然资源语料库的语料资源数据的数据种类及用途详见表A.1。

表 A.1 规划资源语料数据类别

语料信息类别	用途
文本数据	用于文本信息的存储和处理
图像数据	用于图像的存储和处理
视频数据	用于表示和记录动态图像信息，传达视觉内容的变化过程
音频数据	以采样率和位深来表征声音信号

A.2 语料表达

A.2.1 文本数据

文本表征的数据，简称文本数据，是以字符串或字符的形式存储，适用于文本信息的存储和处理。文本数据资源的指标和要求见表A.2。

表 A.2 文本数据的指标和要求

序号	指标项	具体要求
1	类别	详见表3
2	语种	汉语、英语、阿拉伯语、俄语、日语等
3	主题领域	参照《中国分类主题词表》（第二版）中的定义
4	数据资源内容	文本及对应的说明或简介
5	编码格式	UTF-8或GB18030编码（符合《信息技术　中文编码字符集》（GB 18030-2022）的要求）
6	文件格式	TXT、DOC、DOCX、PDF[a]、RTF、HTM、HTML、XML
7	文本验收标准	内容完整、无水印/加密限制、文本可复制编辑
[a]如果PDF文件中的文本是图片形式的，应使用OCR工具将其转换为文字文本。		

注：文本数据指标和要求参考T/SAIAS 015—2024第5.2.1条的规定。

文本数据资源分类说明见表A.3。

表 A.3 文本数据资源分类表

类型	说明
政策法规类	国家及地方发布的法律法规、行政规章、政策文件等
技术规范类	行业标准、技术导则、操作手册、规范性文件等
项目报告类	规划方案、可行性研究报告、环境影响评价报告、竣工验收报告等
学术研究类	学术论文、学位论文、研究报告、文献综述等
会议纪要类	政府会议记录、专家评审意见、研讨会纪要等
其他	特殊用途的文本数据，如公众意见征集、公示文件、历史档案等

A.2.2 图像数据

图像表征的数据，简称图像数据，是以像素矩阵的形式存储，每个像素点包含颜色信息，适用于图像的存储和处理。图像数据资源的指标和要求见表A.4。

表 A.4 图像数据的指标和要求

序号	指标或（和）要求	标准
1	类别	详见下表
2	主题领域	参照《中国分类主题词表》（第二版）中的定义
3	数据资源内容	图像及对应文字说明或图像介绍

表 A.4 图像数据的指标和要求（续）

4	图像编码（文件）	栅格图	格式	BMP	JP(E)G	PNG	GIF	WEBP	Ecw	Tiff
5			位深度	1/4/8/24/32	8/24	8/16/24/32	8	8	8	1/4/8/24/32
6			分辨率	屏幕分辨率不低于1024×768，扫描图像的扫描分辨率不低于72dpi						
7		矢量图	格式	AI	CDR	SVG	EPS	Shp		
8			色彩模式	8/24/32	1/8/24/32	8/24/32	8/24/32	/		
9	图像真实性（如需要）			当图像用于城市规划决策、自然资源评估等重要用途时，应达到《图片真实性鉴定技术规范》(SF/T 0153－2023）中非负面的鉴定结果，保证图像内容的真实可靠，避免因虚假图像导致的决策失误。						
10	图像验收标准			除清晰外，对图像素材的随机抽样中，应有80%包含主体（不含主体的图像素材示例，包括但不限于航拍、延时风景摄影等						

注：图像数据指标和要求参考T/SAIAS 015—2024 第5.2.2条。

图像数据资源分类说明见表A.5。

表 A.5 图像数据资源分类表

类型		说明
绘制类图像	绘画	包括城市规划相关的电脑绘画、线条画、地图（如城市地形图、规划图等）
	绘图	涵盖城市规划中的流程图、结构图、原理图、组织图、方框图、统计图、关联图、分布图等多种类型
	图形	由弧边形、直边形、混合形等几何图形构成的图像，可用于表示城市空间布局的简化模型、抽象概念等
	符标	符号、标号、字符、图标等，用于标识城市中的特定地点、设施或概念，如交通标志、公共设施标志等
	表格	包含数据统计信息，如土地利用分类统计表、城市建设指标表等
拍摄类图像	人物	涉及城市规划与自然资源领域的相关人物，以及与城市文化、历史相关的特定人群图像
	植物	涵盖城市绿化中的花、草、树、叶、农作物、蔬菜等植物图像
	自然	包含山、水、天空、草地、沙漠等自然景观图像，为城市规划中的自然保护区规划、城市生态修复、景观设计等提供自然元素的参考
	建筑	展示城市中的各类建筑，如楼、房、亭、塔、庙等，可用于建筑风格研究、历史建筑保护、城市建筑设计等方面
	场景	生活、会议、体育、节日、舞台、购物等不同场景的城市图像，为城市公共空间规划、活动策划、商业布局等提供场景参考
	大场景	包括卫星遥感影像、航空影像、无人机航拍影像、街景影像等通过专业装置设备在不同复杂环境形成的现状拍摄影像，为城市总体规划、区域规划、专项规划等提供全局或较大范围场景参考

A.2.3 视频数据

视频数据通常用于表示和记录动态图像信息，用于传达视觉内容的变化过程。视频数据资源的指标和要求见表A.6。

表 A.6 视频数据的指标和要求

序号	指标项	具体要求
1	视频类型	详见下表
2	主题领域	参照《广播电视和网络视听节目内容标识标签规范》（GY/T 360-2022）中的“内容类内容特征子类别规范词”
3	数据资源内容	视频及对应文字说明或视频介绍
4	视频分辨率	宜1080P（1920x1080像素）及以上
5	视频帧率	25~30帧/秒
6	单一视频时间长度	10秒及以上
7	视频文件格式	mp4
8	视频验收标准	除清晰外，对视频素材的随机抽样中，应有80%包含主体（不含主体的视频素材示例，包括但不限于航拍、延时风景摄影等）

注：视频数据指标和要求参考T/SAIAS 015—2024 第5.2.4条。

视频数据资源分类说明见表A.7。

表 A.7　视频数据资源分类表

类型	说明
新闻资讯	对城市规划重大政策出台、自然资源开发利用新动态、城市基础设施建设重要进展等近发生的、有价值的资讯进行传播报道的视频内容
纪录片	以真实的城市发展历程、自然资源现状与变迁为表现对象，以实地拍摄的城市景观、自然资源风貌为创作素材，并对素材进行艺术加工，使其兼具现实生活中的真实与审美的特征
专题栏目	以城市规划理念、自然资源保护技术、城市空间利用案例等文化、教育、科学方面的主题或事件为表现内容，题材选具有集中性，融入了专业认知和思考，既符合内容纪实性要求，又具有深入探讨的特点。
专业视频	严谨地传递特定的城市规划专业知识、自然资源评估方法等专业信息，由城市规划师、自然资源领域专家或专业团队制作的非娱乐的高质量视频内容
社交媒体视频	符合社交网络传播特点，如短视频和直播等形式，用于宣传城市规划成果、普及自然资源保护知识等。
教育和培训	通过视频形式进行知识传授和技能训练的媒体内容
其它	不在上述类型所涵盖的范围内的视频内容
变化监测视频	使用无人机、车载等非固定感知设备以及各类固定感知设备如摄像头拍摄的视频，用于自然资源监测、城市治安管理、生态气象预测等城市现代化治理领域

A.2.4　音频数据

声音表征的数据，简称音频数据，是以声音波形的形式存储，通常以采样率和位深来表征声音信号。音频数据资源的指标和要求见表A.8。

表 A.8　音频数据的指标和要求

序号	指标项	具体要求
1	类别	详见下表
2	语种	汉语、英语、阿拉伯语、俄语、日语等
3	主题领域	参照《广播电视和网络视听节目内容标识标签规范》（GY/T 360-2022）中的“内容类内容特征子类别规范词”
4	数据资源内容	音频及对应文字说明或音频介绍
5	音频采样率	不小于44.1KHz
6	通道数	双声道/单声道（由原始资料特性决定）
7	单一音频时间长度	60秒及以上
8	量化精度	不低于16位
9	音频文件格式	WAV
10	音频验收标准	对音频素材的随机抽样中，应有95%满足表6中所述情况

注：音频数据指标和要求参考T/SAIAS 015—2024 第5.2.3条。

音频数据资源分类说明见表A.9。

表 A.9　音频数据资源分类表

类型	说明
实地调查类	外业调查过程中采集的语音记录、现场环境音、访谈录音等
专家评审类	规划方案评审会、技术论证会、专家咨询会等场景下的发言录音
公众参与类	公众意见征集、社区座谈、民意调查过程中的音频记录
会议记录类	行政办公会议、项目汇报会议、政策研讨会议的发言录音
设备采集类	无人机、监测站、传感器等设备采集的配套音频信息
其他	特殊用途的音频数据，如历史口述资料、宣传片配音素材等

附 录 B
（资料性）
规划与自然资源领域知识体系

知识体系是将领域知识按照从宏观到微观、从抽象到具体的逻辑关系进行分层组织的结构体系。以上海为例，规划与自然资源领域的知识体系涵盖国土空间规划、测绘与地理信息、自然资源调查监测评价等16个一级分类，56个二级分类，157个三级分类，详见表B.1。

表 B.1 规划与自然资源领域的语料信息知识体系

一级分类	二级分类	三级分类
国土空间规划	区域规划	长三角
		上海大都市圈及临沪地区
	总体规划	上海市城市总体规划
		浦东新区和各郊区国土空间总体规划
		单元规划
		总体规划监测与动态维护
		总体规划实施管控
	详细规划层次	控制性详细规划
		城市设计
		郊野单元村庄规划
		郊野单元控制性详细规划
		市政管线综合规划
	专项规划	民生保障
		生态环境
		综合交通
		市政基础
		陆海统筹
		创新发展
		空间品质
		重点地区
		历史文化名城保护规划
		历史风貌保护与文化保护传承
		地名管理
		雕塑管理
	城市更新	概念内涵
		政策文件
		更新规划方案成果
		典型案例
		15分钟社区生活圈
测绘与地理信息	行业资质资格与信用管理	管理要求
		外省市测绘组织来沪备案
		测绘作业证
		测绘资质
		注册测绘师
	基础测绘管理	管理要求
		基础测绘
	测绘与地理信息管理	管理要求
		地理空间信息工程
		测绘航空摄影
		工程测量
		互联网地理信息服务
		航空摄影测量与遥感
		地图制图

表 B.1（续）

一级分类	二级分类	三级分类
测绘与地理信息	测绘与地理信息管理	导航电子地图制作
		大地测量
		海洋测绘
		公众宣传
		测绘成果质量检验
	测量标志保护管理	管理要求
		迁建审批
		测量巡查维护
自然资源调查监测评价	基础调查监测评价	年度变更调查
		季度变更调查
	专项调查监测评价	城市国土空间监测
		森林湿地调查监测
		水资源调查监测
不动产和自然资源确权登记	不动产地籍调查	权属调查
		不动产测绘
	自然资源地籍调查	自然资源登记单元的划定
		自然资源登记单元代码编制
		自然资源地籍调查工作内容
	不动产登记	日常登记业务
		行政复议、诉讼
	自然资源确权登记	自然资源登记的程序
		自然资源登记信息
自然资源所有者权益	自然资源资产所有者职责清单	——
	自然资源资产清查统计	自然资源资产清查
	自然资源资产负债表	编制制度
		编制内容
	自然资源资产管理考核评价	考核评价内容
		考核评价要求
		考核评价形式
	国有自然资源资产管理情况专项报告	专项报告的内涵
		专项报告工作要求
		专项报告编制流程
		专项报告工作成果
自然资源开发利用	土地征收	征地补偿安置
		土地征收成片开发
		征地信息公开
		房屋协议置换
		自然资源听证
	土地储备	规划计划
		储备项目
	地价管理	公示地价
		城市地价动态监测
		地价管理政策
	土地供给	土地供应方式
		土地供后监管
	节约集约	建设用地节约集约利用
		开发区节约集约利用
	乡村制度改革	农村振兴综合政策
		农村土地资源配置
		乡村审批制度改革
		设施农用地管理

表 B.1（续）

一级分类	二级分类	三级分类
国土空间用途管制	国土空间用途管制	政策法规
		技术标准
		空间准入
	自然资源年度利用计划	概念与理论基础
		计划编制管理
		计划联动管理
		计划执行管理
国土空间整治与生态修复	国土空间综合整治	全域土地综合整治
		市级土地整治
		土地整理复垦
	国土空间生态修复	生态基底调查
耕地保护监督	耕地保护监督	耕地保护政策
		占补平衡制度实施
城市地质	地质调查	概念内涵
		基础地质
		水文地质
		工程地质
		环境地质
		海岸带地质
		地质资源
	地质安全	地面沉降防治
		地质灾害危险性评估
		突发性地质灾害防治
	浅层地热能	浅层地热能开发利用
	地质资料	地质资料汇交保管利用
规划实施	建筑工程建设项目	政策标准
		审批管理
		重点建筑
		专项工作
	风貌保护建设项目	——
	市政交通建设项目	政策标准研究
		项目审批
		实施深化
		批后管理
		基础设施品质提升
数字城市规划管理	量子城市空间智能创新	顶层设计
		城市空间治理大模型研发
		城市治理时空智能垂类应用
		量子城市实施推进
	信息化管理	信息化规划
		基础信息平台管理一张图
		数据管理
		项目管理
		安全管理
		一网通办
		业务统计
产业用地规划资源管理	总体要求	——
	总规统领	上海大都市圈空间格局
		市域空间布局
		重点项目案例

表 B.1（续）

一级分类	二级分类	三级分类
产业用地规划资源管理	资源统筹	规划要素统筹
		土地要素统筹
		存量盘活更新
	产业服务	三项机制
行政服务	行政审批制度改革	行政审批事项动态管理
	城建档案	——
	信访	——
	政务公开	——
	意见提案办理	——
业务监督	监督目标	——
	监督内容	——
	监督方式	市级专项督察
		建设用地审批“双随机，一公开”抽查
宣传展示	专业技术培训	——
	公众宣传	——
	上海城市空间艺术季	城市文化与空间表达
		城市更新与滨水空间治理
		艺术策展与社会参与
	上海文化和自然遗产日	概念定义
		文化遗产
		自然遗产
		上海特色活动与实践
		上海遗产保护重要机构
		赋能城市更新

附 录 C
（资料性）
规划与自然资源领域知识库

规划与自然资源领域的知识库宜包括原始材料类型、标签体系及更新频率等信息，详见表C.1。

表 C.1 规划与自然资源领域知识库

原始材料类型			标签体系		更新频率（建议）
一级分类	二级分类	三级分类	统一数据标签	个性化数据标签	
学术类	学术著作	学术著作	序号、名称、知识领域（知识点体系）、知识库类别（学科、管理或实务）、发表时间、适用地域（全国、上海）、来源（采集单位）、是否开放至互联网	（无）	年度
	学科教材	学科教材			年度
	培训用书	注册规划师			调整后更新
		注册估价师			
		注册建筑师			
		岩土工程师			
		注册测绘师			
	期刊论文	期刊论文			根据出刊时间更新
管理类	政策法规	法律		文号、发布机构、时效状态（是否废止）	月度
		行政法规			
		部门规章			
		地方性法规			
		地方政府规章			
		国务院和部委规范性文件			
		其他规范性文件			
	技术标准	国家标准		标准号、技术归口单位、批准发布部门、状态（是否废止）、实施日期	月度
		地方标准			
		地方规程			
		行业标准			
		团体标准			
	业务规则	办事指南		索引号、发文字号、发布机构、状态（是否废止）	调整后更新
		重要概念释义		（无）	
		非发文业务规则		（无）	
	管理案例	管理案例		（无）	年度
	审批成果	信息数据		（无）	
		档案成果			年度
实务类	实务案例	科研课题		（无）	年度
		研究报告			
		获奖项目			
		方案征集			
		其他案例			
	技术手册	技术手册			调整后更新
	书籍、期刊	书籍、期刊			根据出刊时间更新
	其他	其他			（无）

参 考 文 献

[1]中国分类主题词表（第二版）
[2]中国图书馆分类法
[3]中国档案分类法
[4]SF/T 0153-2023图片真实性鉴定技术规范
[5]GYT 360-2022广播电视和网络视听节目内容标识标签规范
[6]GB/T 4894-2009信息与文献术语
[7]GB/T 22239-2019信息安全技术网络安全等级保护基本要求
[8]GB/T 22240-2020信息安全技术的网络安全等级保护定级指南
[9]GB/T 35273-2020信息安全技术 个人信息安全规范