ICS 26.0XX.XX
CCS I66

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIAS XXX—2026

国际物流供应链高质量数据集 建设指南

Guidelines for the Construction of International Logistics Supply Chain High-Quality Dataset

2026-XX-XX 发布　　2026-XX-XX 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：厦门供应链数智创新有限公司、上海库帕思科技有限公司、上海人工智能实验室、厦门国贸集团股份有限公司、厦门建发股份有限公司、厦门象屿股份有限公司。

本文件主要起草人：

本文件首次制定。

首期执行单位：

国际物流供应链高质量数据集 建设指南

1 范围

本文件规定了国际物流供应链高质量数据集建设的实施路径、技术方法和过程控制要求，明确了数据采集、清洗、标注、质量评估等建设环节的实施要点。

本文件适用于国际物流供应链高质量数据集建设活动的具体实施与过程管理。

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 41867—2022 信息技术 人工智能 术语

3 术语和定义

GB/T 41867—2022界定的以及下列术语和定义适用于本文件。

3.1

高质量数据集建设 data construction

围绕人工智能模型训练与应用需求，对原始数据开展采集、预处理、清洗、标注、质量评估与版本管理等活动，使其形成结构规范、语义清晰、质量可控、可复用的数据集资产的过程。

3.2

原始数据 raw data

未经系统化处理或仅经过初步采集的国际物流供应链相关数据，具有多模态、非结构化和未标准化等特点。主要包括政策文本、业务记录、轨迹信息、单证影像等内容，尚未经过规范化清洗与标注处理。

3.3

建设工具链 dataset construction toolchain

支撑高质量数据集建设过程中数据采集、预处理、清洗、标注、质量评估等环节的集成化工具集合，具备多模态处理、人机协同和过程追溯等能力。

3.4

数据质量评估 data quality assessment

对建设形成的数据集在完整性、一致性、准确性、适用性等方面进行系统性检查与综合评价，以确认其是否满足高质量数据集相关要求。

4 缩略语

以下缩略语使用于本文件

AI：人工智能 Artificial Intelligence

OCR：Optical Character Recognition，光学字符识别

ASR：Automatic Speech Recognition，自动语音识别

5 概述

5.1 基本目标

国际物流供应链高质量数据集建设的目标，是将来源分散、口径不一的原始数据（如贸易合同与单证数据、提单与运单信息、运输轨迹数据、仓储作业记录、口岸通关与关务监管数据、航线与班列运行信息、运价与成本数据、供应链风险事件数据等）转化为结构规范、语义清晰、可计算、可推理、可复用的高质量数据集资产。

建设成果应实现跨主体、跨环节、跨区域的数据关联与知识显性化，支撑国际物流供应链大模型在保供调度、成本预测、跨境合规、异常预警、多式联运应急响应等典型任务中的训练与推理应用。

5.2 基本原则

5.2.1 面向模型适配原则

数据集建设应服务于人工智能（Artificial Intelligence，AI）模型应用需求，数据结构与语义表达应适配模型学习与推理特征。应避免冗余、歧义和不可解释信息，确保数据表达清晰、边界明确。

5.2.2 场景与任务导向原则

数据集建设应围绕国际物流供应链典型业务场景开展，与实际应用任务保持一致。建设目标应与业务需求对齐，避免脱离应用场景的形式化处理。

5.2.3 人机协同与自动化优化原则

数据集建设宜采用自动化工具与人工审核相结合的方式。重复性处理环节可采用自动化方法实现，关键知识与高风险场景应由行业专家参与审核与确认，确保建设结果的专业性与可靠性。

5.2.4 全流程可追溯原则

建设过程应记录数据来源、处理过程、版本变更及关键操作信息，确保数据建设过程可追溯、可验证、可审计。

5.2.5 安全合规嵌入原则

数据集建设应符合相关法律法规和行业监管要求，在采集、处理、标注和交付等环节嵌入安全与合规控制措施，保障数据处理全过程合法、合规、可控。

5.2.6 全链路一致性原则

数据集建设应围绕国际物流供应链全链路业务结构展开，确保不同业务环节之间的数据在时间表达、空间标识、主键关联和语义定义方面保持一致或具备可映射关系，以支撑跨环节推理与系统分析。

6 高质量数据集建设要求

6.1 建设路径

高质量数据集建设宜遵循“采集—预处理—清洗—标注—入库—应用反馈”的全流程管理路径。不同用途数据（预训练、监督微调、思维链、评测）在建设深度上可有所差异，但应遵循统一的结构规范与质量控制要求。

6.2 数据采集

6.2.1 采集原则

数据采集应围绕国际物流供应链知识体系和典型业务环节开展，覆盖国际贸易规则、跨境运输组织、多式联运调度、仓储作业、口岸通关监管、航线与班列运行、运价与成本变化、供应链风险事件及产业链运行等内容。。采集活动应合法合规，明确数据来源、授权边界及跨境流通限制，确保数据可用于模型训练与推理应用。

6.2.2 数据标识

采集数据应建立统一的数据命名规则与标识体系，宜采用“统一标识码+数据名称”的方式进行管理，确保数据来源可追溯、用途可识别。国际物流供应链数据宜建立跨环节主键关联机制，包括但不限

于：订单号、提单号（B/L）、箱号（Container No.）、运单号（AWB）、报关单号、航次号、车次号、港口作业单号、仓库作业单号等关键标识。

数据集应构建统一标签体系，用于标识业务环节（采销/仓储/运输/通关/结算）、贸易方式、运输方式（海运/空运/铁路/公路）、模态类型及用途类别等信息，支撑后续检索、融合与模型调用。

6.2.3 模态类型

数据采集宜根据业务场景需求，覆盖文本、表格、图像、视频、时序数据及多模态复合数据等类型。多模态数据之间宜建立关联关系，如时间对齐、事件映射或业务主键关联，以增强数据整体表达能力。

6.3 数据预处理

6.3.1 格式规范统一

数据预处理阶段应统一编码格式、时间表达、计量单位、字段命名和元数据结构，确保多源数据在结构层面具备可融合性。不同来源数据在字段口径存在差异的，应通过统一映射规则进行标准化处理。

6.3.2 非结构化数据处理

非结构化数据宜进行基础结构整理，包括文本分段与章节还原、提单/发票/装箱单/报关单等国际贸易单证字段OCR、政策条款结构化、影像元数据提取、视频切片与关键帧提取、语音ASR等，为后续清洗与标注提供结构基础。

6.4 数据清洗

6.4.1 一般要求

数据清洗应识别并处理缺失值、异常值、重复数据及明显错误内容，确保数据内容准确、逻辑一致、表达规范。

6.4.2 逻辑一致性校验

对结构化数据，应校验字段取值合理性、跨字段逻辑关系及时间序列连续性；对文本与政策类数据，应核验关键条款、指标和规则表达的准确性；对图像、视频和时序数据，应保证样本清晰度与数据连续性。

对结构化数据，应校验关键字段的业务逻辑一致性，包括但不限于：

a) 时间一致性：订舱时间、装船时间、到港时间、报关时间等节点顺序应合理；
b) 费用一致性：运费、关税、仓储费等费用项与币种、计费单位匹配；
c) 货物一致性：货物重量、体积、箱型箱量与运输工具载荷匹配；
d) 单证一致性：合同、发票、提单、报关单之间的货物描述、数量、收发货人信息一致。

6.4.3 数据脱敏

涉及企业经营信息、客户信息、合同价格、贸易往来对象、轨迹定位等敏感内容的，应在清洗阶段完成脱敏或匿名化处理。跨境数据应明确脱敏策略与可逆性边界，并保留脱敏记录与处理日志。

6.5 数据标注

6.5.1 标注总体要求

数据标注应围绕模型任务目标开展，标签定义应清晰、边界明确，在不同来源数据间保持语义一致。标注活动宜符合相关数据标注通用标准要求。

6.5.2 标注流程

标注流程宜包括任务设计、样本筛选、实施标注、质量校验与结果修正等环节。关键规则类或复杂推理类样本，宜由行业专家参与审核。

6.5.3 用途差异化标注

不同用途数据在标注深度上可有所差异：

a) 预训练数据可不进行任务型标注，但应完成去噪、结构化与主键关联，保证语义可学习；
b) 监督微调数据应明确输入与目标输出对应关系，如“运价预测输入要素—预测结果”“通关合规问答—标准答案”；
c) 思维链数据应体现推理步骤与依据，如“港口拥堵—改配运输方式—成本影响—交付风险评估”的决策链；
d) 评测数据应提供标准答案与评分依据，覆盖典型边界条件，如贸易条款冲突、政策生效时间变化、异常延误场景等。数据入库与版本管理。

6.6 数据入库与版本管理

6.6.1 数据入库要求

建设完成的数据应按统一数据模型入库管理，形成可直接用于模型训练、推理或评测的数据资产包。入库数据应保留来源信息、业务口径说明、建设过程记录、版本号、责任主体及合规标识，实现可追溯、可交付、可复现。

6.6.2 版本控制

数据集更新应建立版本管理机制，记录更新内容、变更范围及影响说明。不同版本之间应保持可比性与可回溯性。

6.7 应用反馈与持续优化

6.7.1 模型验证反馈

数据集宜通过模型训练与场景应用测试进行验证，重点评估其对如保供调度、价格与成本预测、多式联运应急响应等物流供应链场景任务的支撑效果，并分析模型表现与数据质量之间的关联关系。

6.7.2 持续优化机制

根据模型表现与业务反馈，对数据样本结构、标注策略及规则表达进行优化调整，形成“应用验证—问题识别—数据优化”的持续改进机制。

7 建设工具链

7.1 工具链定义

国际物流供应链数据建设工具链是支撑数据清洗、标注、监控及管理的集成化工具集合，具有以下特性：

a) 全链路覆盖：覆盖从原始数据输入到建设成果输出的全过程；
b) 模块化设计：支持功能模块灵活组合、扩展与定制；
c) 领域适配性：适配国际物流供应链多源异构、多模态数据特性及复杂业务规则要求。

7.2 核心工具组件

7.2.1 多模态协同标注工具

多模态协同标注工具应满足以下要求：

a) 功能要求：支持跨模态数据的关联标注，包括文本—单证影像联动、轨迹—事件时序对齐、多源业务数据协同标注等；

示例：贸易政策文本与适用条件协同标注，运输轨迹与节点事件对齐加工，单证影像与结构化字段联动标注。

b) 标注精度应满足国际物流供应链在业务语义一致性、时序对齐及规则准确性方面的要求，误差应控制在规定范围内。

7.2.2 实时监控与质量管控工具

实时监控与质量管控工具应满足以下要求：

a) 功能要求：动态展示数据建设总量、增量及关键质量指标，支持异常识别与预警提示；

b） 技术基准：数据建设状态与质量监控应实现实时响应与动态同步，确保建设进度与质量全过程可监测、可控制。

7.2.3 数据管理与审计工具

数据管理与审计工具应满足以下要求：

a） 功能要求：支持数据版本管理、加工流程追溯及审计记录管理；

b） 技术基准：应提供可追溯的元数据记录机制，确保数据在建设全生命周期内具备可审计性与合规可查性。

7.2.4 国际物流供应链专用扩展工具

根据典型业务场景可设计专用工具，包括：

a） 政策规则工具：支持国际贸易规则、通关政策及合规条件的结构化处理与规则标注；

b） 风险事件工具：支持供应链风险事件、异常处置过程及因果关系的数据标注与关联分析；

c） 运输节点工具：支持港口作业节点、运输轨迹、班轮航线、班列运行等运输过程数据的结构化处理与事件标注。

7.3 工具链安全规范

工具链安全规范应满足以下要求：

a） 建设工具应符合 GB/T 35274—2023 等相关数据安全要求；

b） 建设成果应嵌入可追溯标识或数字水印，确保数据来源与建设过程可追溯。

8 质量评估

8.1 基本要求

参考 ISO/IEC 5259-4:2024 等国际标准，建立系统化的质量评估指标体系与流程。

根据标准框架，数据质量可从多个维度进行评估，国际物流供应链数据具有以下特点：

a） 多源异构与多模态特性：涉及政策文本、业务记录、单证影像、轨迹数据、时序数据等多种形态；

b） 业务决策导向：数据主要服务于贸易合规判断、运输组织、风险预警、应急响应等核心业务场景；

c） 获取与建设成本高：部分数据涉及跨境合规、商业敏感与实时性要求，需优先保障业务可用性与决策可靠性。

本指南在全面参考相关标准的基础上，优先选取准确性、一致性、适用性三个维度作为核心质量评估指标，覆盖数据质量的基本要求，并支撑国际物流供应链典型应用场景需求。

8.2 准确性

准确性指数据内容正确反映其所描述的真实业务对象、状态或事件的程度。评估内容与方法包括：

a） 结构化与时序数据：与来源系统或权威数据进行比对；检查数值是否超出合理业务范围；利用多源数据进行交叉验证；

b） 图像与视频数据：人工抽样检查关键字段或标注是否正确；验证单证影像与对应结构化字段的一致性；

c） 文本与政策类数据：人工审核关键条款、条件与结论的准确性；与权威政策发布渠道进行比对；

d） 标注数据：通过人工质检抽查评估标注准确率、精确率、召回率、F1 值等指标；

e） 业务逻辑一致性：检查数据项之间是否符合国际物流供应链业务规则与逻辑约束。

8.3 一致性

一致性指数据在不同来源、不同时间或不同表达形式下保持无矛盾状态的程度。评估内容与方法包括：

a） 时间一致性：检查关联数据的时间戳是否匹配、顺序是否合理；

b） 格式一致性：检查数据格式、编码规则、计量单位是否遵循统一规范；
c） 值域一致性：检查数据取值是否处于定义的合理范围或合法枚举集合内；
d） 关联一致性：检查关联引用的数据是否存在、可访问且匹配；
e） 标注一致性：评估不同标注人员对相同或相似数据的标注结果一致性。

8.4 适用性

适用性指数据满足特定国际物流供应链应用场景需求的程度，评估内容与方法包括：
a） 相关性：评估数据是否覆盖目标业务场景的关键要素与核心变量；
b） 知识密度：评估数据中蕴含的规则信息、经验知识与决策逻辑价值；
c） 多样性：评估数据是否覆盖足够的业务类型、运输方式、贸易条件与风险情形，以支撑模型泛化能力；
d） 时效性：评估数据更新频率与新鲜度是否满足业务应用需求；
e） 行业特性覆盖性：评估数据是否覆盖港口、航运、铁路班列、航空货运、仓储节点、跨境通关等国际物流关键环节

9 安全与合规控制

9.1 数据脱敏

数据脱敏应遵循“最小必要”原则：个人身份信息应实施去标识化处理；企业商业秘密应采用泛化或屏蔽处理；国家敏感信息应严格执行保密规定。脱敏规则应经安全专家评审，脱敏效果应定期验证。

9.2 访问控制

建设过程应实施分级访问控制，根据数据敏感级别设置差异化操作权限。关键操作应实行双人复核机制，敏感数据访问应履行审批程序并留存操作痕迹，定期开展访问日志审计以识别异常行为。

9.3 审计与留痕

应完整记录数据来源与授权信息、各环节操作执行人与时间戳、处理参数与质量检查结果、问题整改过程等关键信息。应保留原始数据与各阶段版本，支持问题回溯与责任界定，审计日志保存期限不应少于三年。

参 考 文 献

[1] 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
[2] GB/T 35274-2023 数据安全技术 大数据服务安全能力要求
[3] 中国档案分类法
[4] 中国分类主题词表（第二版）
[5] 中国图书馆分类法