多模态公共数据集授权运营体系的构建与实施路径研究

甄理[1]，陈龙秋枫[2]，章昱昕[2]，俞佳敏[3]，王亚妮[3]
1. 浙江省数据局省大数据发展中心，浙江 杭州 310000
2. 浙江省可信数据要素研究院，浙江 杭州 310000
3. 中电云计算技术有限公司，浙江 杭州 310000
DOI: 10.61369/TACS.2026090003

摘　　要： 公共数据资源是数字时代国家基础性战略资源，其开发利用水平深刻影响着国家数字化转型与人工智能产业发展进程。本文系统分析多模态公共数据集建设的政策基础、现实需求与核心痛点，提出以全生命周期“一本账”统筹五维协同建设的总体路径，构建“原始数据层—基础层—领域层—场景层”四级分层架构与“不敏感—低敏感—较敏感—敏感”四级安全分级体系。在此基础上，贯通需求识别、治理加工和模型验证等八个环节的建设路径，设计“一底座、三平台”的技术架构与权责清晰、分类授权的运营模式，并建立全流程安全治理与市场化收益反哺机制。研究表明，通过分层分类、安全可控、价值闭环的体系化设计，能够有效破解多模态公共数据“供数动力不足、流通效率不高、收益分配不清”等核心难题，为数据要素市场化配置改革提供实践参考。

关 键 词： 多模态公共数据；数据集建设；授权运营；数据要素；人工智能

Research on the Construction and Implementation Path of the Authorization Operation System for Multimodal Public Datasets

Zhen Li [1], Chen Longqiufeng [2], Zhang Yuxin [2], Yu Jiamin [3], Wang Yani [3]
1. Zhejiang Provincial Big Data Development Center, Zhejiang Provincial Data Bureau, Hangzhou, Zhejiang 310000
2. Zhejiang Institute of Trustworthy Data Elements, Hangzhou, Zhejiang 310000
3. China Electronics Cloud Computing Technology Co., LTD., Hangzhou, Zhejiang 310000

Abstract: Public data resources are fundamental strategic resources for a country in the digital age. The level of their development and utilization profoundly influences the process of a country's digital transformation and the development of the artificial intelligence industry. This paper systematically analyzes the policy basis, practical demands and core pain points of multimodal public dataset construction, proposes an overall path of five-dimensional collaborative construction coordinated by a "single account" throughout the entire life cycle, and constructs a four-level hierarchical architecture of "raw data layer - basic layer - domain layer - scenario layer" and a four-level security classification system of "insensitive - low-sensitive - relatively sensitive - sensitive". On this basis, the construction path of eight links including demand identification, governance processing and model verification is integrated. A technical architecture of "one base and three platforms" and an operation mode with clear rights and responsibilities and classified authorization are designed, and a full-process security governance and market-based revenue feedback mechanism are established. Research shows that through a systematic design featuring hierarchical classification, security controllability, and value closed-loop, the core problems of multimodal public data such as "insufficient supply power, low circulation efficiency, and unclear revenue distribution" can be effectively solved, providing practical references for the market-oriented allocation reform of data elements.

Keywords: multimodal public data; dataset construction; authorized operation; data elements; artificial intelligence

引言

公共数据资源是数字时代国家正常运行和变革创新发展的基础性、全局性、战略性资源，其开发利用广度、深度和水平，深刻影响着国家数字化转型和现代化建设进程。近年来，国家密集出台“数据二十条”[1]、《关于加快公共数据资源开发利用的意见》[2]等政策文件，从数据产权、流通交易、收益分配、安全治理四方面搭建数据基础制度体系，明确提出推进实施公共数据确权授权机制，以“使用

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

权”为核心的权利束体系为公共数据授权运营奠定顶层框架。2026年，国家数据局发布《关于推进行业高质量数据集建设行动的实施方案（征求意见稿）》[4]，聚焦人工智能预训练、指令微调等各阶段数据需求，推动构建符合结构完整性、内容多样性、标注准确性、模型适配性等质量标准的高质量数据集。

浙江省作为全国数字化改革先行省份，率先构建起较为完善的公共数据治理与授权运营政策体系[7]。《浙江省公共数据资源授权运营管理办法》[5]明确由各级数据主管部门作为实施机构，采用“依场景授权”模式，依托一体化公共数据平台或可信数据空间建设授权运营域，确保“原始数据不出域、数据可用不可见”。然而，当前公共数据授权运营主要依托结构化数据开展，面向人工智能大模型与智能体应用的文本、图像、音频、视频等多模态公共数据授权运营仍处于起步探索阶段，面临智算基础设施支撑不足、专业化治理工具欠缺、标准体系缺失、运营机制待完善等现实挑战。在此背景下，系统构建多模态公共数据集授权运营体系，探索可落地、可复制、可推广的建设路径，对赋能人工智能产业高质量发展具有重要的理论价值与实践意义。

一、建设基础与现实需求

（一）政策基础与建设现状

国家层面，“数据二十条”提出推进公共数据确权授权机制；《公共数据资源授权运营实施规范（试行）》[3]细化授权运营全流程要求，明确“整体授权、分领域授权、依场景授权”等模式；《数据安全法》《个人信息保护法》确立数据安全与隐私保护底线。省级层面，浙江省《公共数据资源授权运营管理办法》[5]实现从“试行探索”向“全面实施”的升级，明确场景申报、方案评审、数源审核、协议签订等标准化流程；《关于支持人工智能创新发展的若干措施》[6]明确提出加强高质量公共数据供给。

浙江省依托一体化智能化公共数据平台，已建成较为完善的授权运营域，落地100余个授权运营场景，在数据归集、目录管理、申请授权、安全管控等方面形成可复制模式。然而，目前授权运营主要依托结构化数据开展，多模态数据开发仍处于起步阶段。医疗领域“安诊儿”等场景虽已形成重要实践样本，验证了多模态数据赋能AI服务的应用价值，但面向大模型训练与智能体开发的多模态公共数据集建设，在智算支撑、治理工具、标准规范、运营机制等方面仍存在明显短板。

（二）核心痛点分析

当前浙江省多模态公共数据集建设面临六方面关键瓶颈：一是数据需求不清晰，尚未形成系统化需求分析体系和供需对接机制，导致数据供给与产业需求错配；二是数据供给能力不足，文本、图像、音频、视频等多模态非结构化数据尚未实现系统化归集和规模化供给；三是基础设施和治理平台支撑不足，现有设施主要面向传统数据加工场景，在智算资源、多模态治理、模型训练等方面能力薄弱；四是安全治理体系有待完善，针对多模态数据的分类分级、脱敏处理、模型安全治理等制度规范仍不健全；五是运营服务生态协同不足，尚未形成覆盖数据治理、标注、训练、评估等环节的专业化服务体系；六是评估与价值反馈机制不健全，缺乏统一的质量评价标准和运营成效评估体系，难以形成“建设—运营—评估—优化”的持续改进闭环。

二、总体实施路径与分层体系

（一）总体路径：一本账统筹与五维协同

本文提出以建设公共数据集全生命周期“一本账”为总抓手，统筹推进资源编目、分层建设、运营管理和价值释放，实现数据资源底数清晰、建设过程可追溯、运营情况可监管、安全风险可防控、应用成效可评估。在此基础上，构建“政策规划—平台建设—分层建设—安全兜底—价值闭环”的五维实施思路：政策规划明确制度边界与发展方向；平台建设夯实技术底座与支撑能力；分层建设推动公共数据集体系化开发与目录化管理；安全兜底保障数据全生命周期安全可控；价值闭环促进数据资源持续优化与价值释放。

（二）分层分类建设体系

借鉴数据仓库分层治理思想，将公共数据集划分为四级架构：原始数据层（D0）包括从IRS平台编目数据和其他平台汇聚的未加工数据，原则上仅在授权运营域内供预处理使用；基础层（D1）聚焦社会信息、工商信息、自然资源等通用高频公共基础数据集，为通用大模型训练提供基础语料；领域层（D2）面向医疗健康、工业制造、交通运输等重点行业，构建行业通识数据集；场景层（D3）针对疾病筛查、金融风控、设备故障预警等具体AI应用，开发场景化专业数据集。

同时，根据数据敏感程度将公共数据分为S1（不敏感）、S2（低敏感）、S3（较敏感）、S4（敏感）四级。S1级数据可无条件开放；S2级数据在脱敏或取得特定主体同意后受限开放；S3级数据须经审批后在授权运营域内加工使用；S4级数据原则上禁止出域，仅在严格隔离环境中供特定主体在监管下使用。不同敏感等级与不同数据层级的交叉组合，决定了数据在加工、训练、推理、出域等环节的权限边界与流转规则。

（三）实施原则与阶段目标

公共数据集建设遵循四项基本原则：可操作性原则，确保各项路径、机制与措施具备落地实施条件；可评价性原则，建立覆盖建设质量、运营成效、安全合规的多维评估体系；可推广性原则，在试点验证基础上凝练形成“浙江模式”；安全合规原则，始终将数据安全与个人隐私保护置于首位。

按照试点验证、规模扩展和生态成熟分步推进：第一阶段（2026年）完成方案设计与试点验证，在医疗健康、交通物流等领域建成1-2个重点领域数据集体系；第二阶段（2027年）实现规模扩展，将验证模式复制到3-5个重点领域，初步形成市场反馈驱动数据供给优化的良性互动格局；第三阶段（2028年）实现生态成熟，公共数据集体系全面赋能全省人工智能产业，凝练成可

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

复制、可推广的完整方法论。

三、多模态数据集全生命周期建设路径

依据《高质量数据集建设指南》[8]，结合浙江省公共数据资源授权运营工作实际，将数据集建设划分为需求分析、数据规划、数据采集、数据脱敏、数据清洗、数据标注、数据质检、模型验证八个核心环节，形成闭环优化的建设流程。

（一）需求分析与数据规划

需求分析以人工智能应用场景牵引数据集建设方向。围绕通用大模型、垂直领域大模型、智能体三类应用形态，明确各领域行业分布与核心应用场景；区分不同应用对数据类型、量级、格式的差异化需求；通过“企业深度访谈 + 场景案例分析”形成“领域—场景—需求占比”对应关系，清晰界定需求优先级。数据规划按照 D0-D3四级体系，明确各层级覆盖范围、数据模态组合、存储格式与接口标准；统一调研数据质量标准，要求结构化数据关键字段缺失率、错误率满足阈值，多模态数据标注准确率、格式兼容性达到质量要求。

（二）数据采集与差异化脱敏

数据采集建立覆盖存量、规模和时效的归集机制。全面梳理传统结构化数据与多模态非结构化数据的存量规模，建立“领域—数据类型—存量—质量”台账；结合模型训练需求与成本约束，充分考虑数据冗余率、清洗损耗率、标注有效率等因素预留裕量；明确数据更新频率与模型训练周期的匹配关系，建立动态更新机制。

数据脱敏针对不同模态实施差异化安全处理。结构化数据对姓名、身份证号、联系方式等个人敏感字段及企业敏感信息进行掩码、泛化或加密；图像视频中的人脸、车牌、证件等敏感区域进行模糊化、遮挡或替换；音频中的声纹特征进行变声或降噪处理；文本中的实体信息进行识别与替换。采用“规则脱敏 +AI 脱敏”双轨模式，脱敏后通过可逆性检测确保数据不可恢复原样。

（三）清洗、标注与质检

数据清洗提升数据规范性、完整性和可用性。结构化数据实施缺失值处理、异常值检测、重复值去重、逻辑一致性校验；非结构化文本进行编码统一、乱码过滤、敏感词过滤；图像进行分辨率归一化、色彩空间统一、损坏剔除；视频进行关键帧提取、音画同步校验；音频进行采样率统一、噪声抑制。建立清洗前后数据质量对比机制，量化清洗增益。

数据标注强化语义赋能与跨模态关联。文本采用实体标注、关系标注、情感标注；图像采用分类标注、检测框标注、语义分割；视频采用时序行为标注、目标跟踪；音频采用语音转写、声纹识别标注。实施“初标—复核—抽检”三级质控，关键领域引入专家仲裁。多模态数据对齐是构建高质量数据集的关键，需建立语义对齐、时序对齐、空间对齐机制，确保同一实例在不同模态下的表征一致性。

数据质检统筹数据质量、标注质量与脱敏效果核验。数据质量设置完整性、准确性、时效性、一致性四项核心指标；标注质量设置准确率、一致性、完备性三项指标；部署基于规则引擎和机器学习模型的自动化质检工具。脱敏质检作为独立安全环节，通过可逆性检测、残留风险检测、合规性审查，确保脱敏数据不可恢复原样，未通过数据一律退回重新处理。

（四）模型验证与迭代优化

模型验证以应用效果反馈驱动数据集持续迭代。训练验证监控损失曲线、收敛速度与过拟合情况；性能验证在独立测试集上评估模型精度、召回率、F1值等任务指标；场景验证在真实业务环境中评估模型鲁棒性与业务价值；将验证结果反馈至建设前端，对数据缺口、标注偏差、分布不均等问题启动针对性增补，形成“建设—验证—优化”闭环。

四、平台架构与授权运营机制

（一）“一底座、三平台”技术架构

为支撑多模态数据从归集、脱敏、清洗、标注、质检、融合到模型训练、智能体开发的全流程建设，构建“一底座、三平台”的多模态公共数据授权运营平台。

底座能力集成算力、存储、资源调度和可信空间联通能力。算力方面构建通用计算（CPU 集群）与智能计算（GPU 集群）的层次化体系，分别支撑数据集基础治理与大模型训练；存储方面构建对象存储、向量数据库、关系型数据库三类核心能力，满足非结构化数据归档、语义检索与元数据管理需求；资源运营管理实现多租户隔离、异构资源统一调度与计量计费；通过标准化连接器实现授权运营域与省级可信数据空间的“统一身份、统一标识、统一目录”联通，实现数据集出域的字段级权限控制、脱敏效果校验与智能体服务出域的 API 调用管控、内容过滤。

平台工具包括多模态数据治理平台、模型开发平台、智能体应用开发平台。数据治理平台为文本、图像、音频、视频提供接入、解析、处理、增强、质检全流程工具，支持政务公文、医学影像、监控视频、热线音频等差异化治理，以“目录 + 卷 + 文件”三层资产结构实现数据集版本控制、血缘追踪与一本账管理。模型开发平台面向领域大模型精调与场景模型训练，提供交互式建模、分布式训练、模型评估与安全检测。智能体应用开发平台提供可视化编排、知识库挂载、RAG 检索增强与 API 服务发布，支撑行业智能体快速落地。

（二）授权运营机制设计

公共数据集授权运营涉及实施机构、运营机构、数源单位和使用方等多方主体，需建立权责清晰、运转高效的运营模式。

主体职责方面，实施机构统筹编制建设总体方案，组织场景申报评审与协议签订，实施全流程监管；运营机构在授权范围内加工处理数据，形成数据产品和服务，接受年度评估；数源单位负责数据资源目录编制、需求审核与数据质量保障；使用方按协议约定合规使用数据产品，不得超范围使用。

授权策略实施分层分类管理。D0层数据原则上不直接开放；D1基础数据集采用“整体授权 + 依场景授权”，以基础使用费 + 按调用量计费定价；D2领域数据集采用“分领域授权”，以领域

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

专有权费 + 增值分成定价，同一基础数据集可支撑多个领域数据集，使用方无需重复支付基础成本；D3场景数据集采用“依场景授权”，以场景授权费 + 收益分成定价，授权周期与场景生命周期绑定。S1级数据授权流程简化，S2级需强化审核与脱敏，S3/S4级原则上不出域，仅通过隐私计算技术实现“可用不可见”。

市场运营完善定价、复用、成本分摊和收益反哺机制。构建覆盖数据层级与建设环节的差异化定价体系，D0/D1层以成本补偿为主，D2层增加行业知识溢价，D3层实行市场议价。建立“谁受益、谁付费”的成本分摊机制与数据复用机制，基于同一基础数据集构建高阶数据集时仅需支付增量加工费用。收益分配方面，数源单位获得基础使用费的30%-40%，运营机构获得加工服务费与增值分成的40%-50%，平台管理方获得10%-20%运维费，形成“成本补偿 + 增值分成 + 政策补贴”的激励闭环。

五、全流程安全治理与保障机制

（一）风险识别与技术防护

多模态公共数据在归集、加工、训练、流通和应用环节面临敏感信息泄露、脱敏不彻底、数据出域风险、人工审核滞后等安全风险。构建“事前防接触、事中强管控、事后可追溯”的技术防护体系：原始数据防接触控制实施抽样与前置脱敏，运营机构全过程不接触原始数据；数据内容安全治理建立覆盖文本、图像、音频、视频的统一敏感信息识别与脱敏规则库，实现敏感信息全流程智能化管理；计算存储与隔离防护实行网络隔离、访问控制与加密存储；数据分级管理依据S1-S4等级在加工、向量化、服务调用全过程执行差异化安全策略；数据围栏与出域防护对API出参、向量检索结果、模型输出内容实施强制扫描与二次脱敏；智能合规与全程审计将合规检查结果同步归入高质量数据集“一本账”，实现安全管控与台账管理联动闭环。

（二）制度监管与保障机制

制度监管方面，健全事前审批、事中监测、事后审计和应急处置机制。制定公共数据分类分级、多模态数据脱敏、数据出域管理等配套制度；建立闭环审核监管机制，严格执行抽样方案审批、脱敏策略备案、数据出域审核；压实运营机构安全主体责任，实行安全负责人制度；建立日常监测、定期检查、专项评估、第三方审计相结合的监管模式。

实施保障从政策、资金、技术、生态四个维度建立全面保障体系。政策支持方面，制定人工智能公共数据集建设实施细则、质量评估管理办法等配套文件，优化审批流程，探索“语料券”“算力券”等激励政策；资金投入方面，设立人工智能数据要素发展基金，通过PPP、特许经营等方式引入社会资本，建立“谁受益、谁付费”的成本分摊机制；技术支撑方面，升级一体化智能化公共数据平台与可信数据空间，开展多模态数据融合、小样本标注、隐私计算等关键技术攻关，制定地方标准规范；运营生态方面，建立服务商分级认证体系，构建“原始数据—加工数据集—数据产品”完整加工链条，培育数据交易市场与专业人才队伍。

六、重点领域试点与推广

为验证建设路径与运营模式的可行性，选取医疗健康、交通运输两个领域作为首批试点。医疗健康领域重点构建医疗影像数据集、电子病历数据集、医保结算数据集，验证高敏感数据的安全流通经验；交通运输领域重点构建智能交通视频数据集、物流运输数据集、港口作业数据集，验证多模态数据融合技术路径。

试点实施分三阶段推进：第一阶段完成方案设计与资源准备，成立试点工作组，完成数据资源盘点与需求调研；第二阶段开展数据集建设与产品开发，按照八大环节完成数据采集、脱敏、清洗、标注、质检与模型验证；第三阶段完成授权运营与成效评估，签订授权运营协议，完成产品出域审核与上架，形成可复制推广的标准模板。通过“定量评估 + 定性评估”相结合的方式，从建设质量、应用效果、安全合规、经济效益等维度建立评估指标体系，对成效显著的试点场景进行标杆示范，逐步向工业制造、金融服务、文化旅游等领域复制推广。

七、结语

多模态公共数据集建设是人工智能产业发展的关键基础设施，其供给水平直接决定模型性能上限。本文以浙江省为例，系统构建了“一本账”统筹、五维协同、分层分类的公共数据集建设体系，贯通需求分析、数据采集、脱敏清洗、标注质检、模型验证等全生命周期环节，设计“一底座、三平台”的技术架构与权责清晰、分类授权的运营模式，建立全流程安全治理与市场化收益反哺机制。研究表明，通过D0-D3四级分层与S1-S4四级分级的交叉管理，能够有效平衡数据安全与价值释放；通过“成本补偿 + 增值分成 + 政策补贴”的差异化收益分配，能够破解供数动力不足与收益分配不清的难题；通过医疗健康、交通运输等重点领域试点，能够以“小切口、大深纵”的实践验证路径可行性，为全省规模化建设积累经验。

浙江省作为全国数字化改革先行省份，有条件、有能力在公共数据集建设领域率先突破。未来应进一步加快标准完善、平台建设、试点验证和生态培育，推动公共数据集体系全面赋能人工智能产业高质量发展，为全国数据要素市场化配置改革贡献“浙江智慧”。

参考文献

[1] 中共中央，国务院 . 关于构建数据基础制度更好发挥数据要素作用的意见 [Z].2022.
[2] 中共中央办公厅，国务院办公厅 . 关于加快公共数据资源开发利用的意见 [Z].2024.
[3] 国家发展改革委，国家数据局 . 公共数据资源授权运营实施规范（试行）[Z]. 发改数据规〔2025〕27号 ,2025.
[4] 国家数据局 . 关于推进行业高质量数据集建设行动的实施方案（征求意见稿）[Z].2026.
[5] 浙江省人民政府 . 浙江省公共数据资源授权运营管理办法 [Z].2025.
[6] 浙江省人民政府 . 关于支持人工智能创新发展的若干措施 [Z].2025.
[7] 浙江省数据局 . 浙江省公共数据资源授权运营工作指引 [Z].2025.
[8] 全国数据标准化技术委员会 . 高质量数据集建设指南 :TC609-5-2025-01[S].2025.

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License