

行业高质量数据集链主模式构建与水运行业实践研究

王惠¹ 白晶² 董晓岩²

(1. 大连东软信息学院

2. 交通运输部水运科学研究院)

DOI: 10.65285/QQJJGLLT.V2I2.006

摘要: 数据要素已成为驱动新质生产力发展的核心生产资料, 行业高质量数据集是人工智能模型训练、产业数字化转型的底层数字底座。链主作为行业高质量数据集建设的核心主体, 承担资源统筹、生态组织、全产业链协同引领三重职能。本文以水运行业为研究载体, 围绕高质量数据集图谱搭建、特色数据集开发、全流程管理体系、场景化应用体系、全生命周期技术支撑体系五大维度, 系统阐释行业高质量数据集链主模式完整框架, 为交通行业数据要素市场化配置、以数据赋能产业高质量发展提供实践范式与理论参考。

关键词: 数据要素; 高质量数据集; 链主模式; 水运行业; 数据治理; 场景赋能

Research on the Construction of Chain Leader Model for Industrial High-Quality Datasets and Its Practice in Water Transport Industry

Hui Wang¹, Jing Bai², Xiaoyan Dong²

1. Dalian Neusoft University of Information

2. China Waterborne Transport Research Institute, Ministry of Transport

Abstract: Data elements serve as the core production materials driving the development of new quality productive forces, and industrial high-quality datasets act as the fundamental digital base for AI model training and industrial digital transformation. Different from the standardized development of single datasets, the construction of industrial datasets integrates technical engineering with ecological coordination and value reconstruction. As the core subject of high-quality dataset construction, chain leaders undertake three major functions: resource coordination, ecological organization and whole industrial chain collaboration guidance. Taking the water transport industry as the research carrier, this paper systematically constructs a complete framework of the chain leader model from five dimensions: industrial dataset graph construction, characteristic dataset development, full-process management system, scenario-based application system and full-life-cycle technical support system. It resolves prominent problems including data fragmentation, insufficient collaboration efficiency, uneven data quality and disorderly data circulation in water transport. The research provides theoretical references and practical paradigms for the market-oriented allocation of data factors and data-enabled high-quality industrial development in the transportation sector.

Keywords: data elements; high-quality datasets; chain leader model; water transport industry; data governance; scenario empowerment

前言

数据要素市场化配置是培育新质生产力、推动产业转型升级的核心抓手^[1]。高质量数据集作为 AI 模型训练、行业智能决策、风险预警、绿色低碳发展的基础资源, 其建设水平直接决定产业数字化转型深度。当前各行业普遍面临数据资源碎片化、跨主体协同不足、数据质量管控缺失、数据流通合规机制不完善、场景落地与数据供给脱节等共性难题。传统单一企业独立建库模式存在明显局限性: 中小企业缺乏数据治理技术与资金投入, 大型企业数据封闭形成数据孤岛, 行业整体陷入“数据割据”零和博弈, 难以释放数据要素乘数效应。在此背景下, 依托行业龙头、科研机构、头部平台形成“链主”引领模式, 统筹全行业数据资源共建共享, 成为破解行业数据发展困境的可行路径^[2]。

本文立足水运行业实践, 系统剖析行业高质量数据集链主模式内在逻辑、实施路径、体系架构, 厘清链主在数据资源整合、治理能力输出、场景落地、生态共建中的核心职能, 为交通及其他行业数据要素市场化落地、数字产业生态构建提供理论与实践支撑。

一、行业高质量数据集与链主模式核心内涵

(一) 行业高质量数据集内涵与建设逻辑

行业高质量数据集是经过标准化采集、清洗、标注、校验全流程处理, 适配行业业务场景、可直接用于人工智能模型训练与产业智能分析的标准化数据集。单一数据集建设遵循“数据需求-数据规划-数据采集-数据预处理-数据标注-模型验证”线性流程, 但面向全产业链的行业级数据集建设远超单一数据集开发范畴, 兼具技术工程与

战略生态双重属性。高质量数据集的核心价值在于打通数据流通堵点,推动行业数据应用从分散、盲目、碎片化的“盲人摸象”式探索,转向依托完整数据资源图谱开展精准化、规模化、体系化的科学共建,充分释放数据要素赋能产业升级的价值潜力^[3]。

(二) 链主模式的定位与核心职能

行业高质量数据集建设链主是统筹全行业数据集开发运营的核心支撑主体,一般由行业科研院所、头部龙头企业、国家级行业平台承担,具备三大核心定位:

第一,行业数据集建设领跑者。链主具备完整的数据技术能力、行业业务认知与资源统筹权限,牵头开展全行业数据资源普查、标准制定、基础平台搭建,承担高投入、长周期的基础性数据治理工作,弥补中小市场主体技术与资金短板。

第二,全链条建设组织者。链主搭建行业数据共建共享联盟,统筹监管机构、供需企业、技术服务商等多方主体,统一数据采集、标注、质量、流通标准,建立数据争议处置、收益分配、合规审查配套制度,理顺多方协同机制。

第三,产业数据生态引领者。链主摒弃封闭化技术壁垒思维,将成熟的数据治理能力模块化、开源化输出,打造即插即用的数据服务组件,推动行业数据治理经验共享,打破数据孤岛与数据割据,构建共赢型数字产业生态。

链主模式区别于企业自建数据体系的核心优势在于:跳出单一主体利益局限,以全行业高质量发展为目标,统筹数据资源、技术能力、业务场景、制度规则四大要素,形成“链主引领、多方参与、标准统一、价值共享”的可持续数据集运营体系。

二、水运行业高质量数据集构建体系

数据集图谱是行业数据资源管理的底层载体,以“数据集”为核心节点,依托语义关系关联数据来源、属性、应用场景、业务任务,构建可视化知识网络,实现全行业数据资源全景呈现。

(一) 图谱三元组分层架构

图谱采用“实体-关系-属性”三元组基础架构,分层完成全产业链要素覆盖。实体层划分五大类别:基础设施类(港口、航道、码头、航标、锚地)、运输装备类(船舶、集装箱、装卸设备)、运输服务类(航线、货运订单)、行业主体类(航运企业、港口运营商、船代、海事监管机构)、动态要素类(运量、运价、船舶密度、水文气象、安全事故、风险隐患);关系层构建四类关联逻辑:空间关系、业务流转关系、动态时序关系、时间演化关系;属性层定义三类标准信息:技术属性、业务运营属性、行业监管管理属性。

(二) 图谱分阶段落地实施

水运数据集图谱工程规模大、周期长,分三阶段稳步推进:

第一阶段:基础数据归集与框架搭建。整合现有存量数据资源,开展水运全链条数据普查,联合港口、航运、海事部门采集运营、航行、货运、监管原始数据,完成清洗、

标准化转换,搭建图谱基础框架,实现基础数据可视化展示。

第二阶段:多源数据深度融合与功能优化。统一行业数据标准与编码体系,迭代图谱模型结构,开发智能检索、关联分析等AI功能,完善跨场景数据融合能力。

第三阶段:全产业链成熟图谱体系建成。形成覆盖水运全链条、实时动态更新、具备智能分析能力的完整数据集图谱,面向行业内外机构提供一站式数据检索、调用、分析服务。

三、水运行业高质量数据集标准化管理体系

数据治理是行业数据集建设的核心难点,链主牵头搭建覆盖数据全生命周期的标准化管理体系,从顶层原则、质量管控、合规争议处置、多模态清洗引擎、轻量化服务输出五大维度形成完整治理闭环。

(一) 体系建设总体原则

管理体系以“航道-港口-船舶-物流”为主数据链条,确立“智能化标注、多模态融合、规则约束、AI驱动”核心治理路径,分为管理、技术双支撑维度^[4]:

管理维度:建立覆盖采集、清洗、标注、评测全流程标准化质量管理体系,明确数据授权使用、争议处置、内容合规审查规则;

技术维度:依托多模态清洗引擎、智能质量评估模型、区块链可信存证技术,打造“标准化治理框架+质量闭环体系”双核架构。

体系建设目标是打通原始数据到业务可用数据的无损转化通道,实现结构化、半结构化、非结构化全域水运数据一体化治理。

(二) 全流程数据质量管理闭环

构建“规则刚性约束+AI柔性治理”双模式质量管控体系,搭建三级质量评估指标,保障数据完整性、一致性、时效性。

第一,规则引擎构建业务冲突检测库。设置时序矛盾、空间越界、数值超限硬性规则,实时拦截问题数据;融合生成式AI识别AIS轨迹漂移、设备异常停机、货运量突变等非常规异常,弥补固定规则局限性。

第二,动态量化质量评分机制。基于贝叶斯网络计算数据健康度,生成可视化质量热力图;区块链全程存证质检、修复全流程日志,支持数据质量全链路溯源审计。

第三,评估-修复-优化闭环联动。质量报告反向驱动采集接口升级、清洗规则迭代,持续优化数据源头质量。

同时,链主牵头制定四项行业统一规范:《水运行业数据采集操作指南》《水运行业数据清洗质量评估指标》《水运行业数据标注质量控制标准》《水运行业数据质量测评管理规范》,建立“初标-复检-抽检”三级人工复核流程,定期发布行业数据质量评估报告,动态迭代治理策略^[5]。

四、水运行业高质量数据集场景化应用体系

数据价值释放的核心路径是数据与业务场景深度融合,将数据要素嵌入国家战略、区域发展、行业治理、企业运营全层级,构建“场景牵引、以用促建”的应用体系,搭

建分层分类示范场景,打通需求与供给双向联通渠道。

(一) 分层级高价值应用场景构建

按照宏观、中观、微观三层维度设计差异化示范场景,覆盖战略落地、产业协同、企业精细化运营全层级需求。

1. 宏观场景: 国家战略创新标杆

宏观场景立足全国产业发展全局,服务中长期水运发展规划,聚焦沿海港口基建、内河航道联通、双碳绿色转型、行业现代化治理等战略目标。典型特征为跨产业协同、长期战略价值、政策市场双向驱动,例如交能融合、全域绿色港口建设、全国铁水联运一体化调度。

宏观场景数据融合重点解决大规模多源异构数据全生命周期管理、敏感数据安全保障难题,依托顶层设计、统一数据规划、全行业生态共建,以数据驱动产业格局重构。

2. 中观场景: 区域协同应用示范

中观场景对接京津冀、长江经济带、长三角、粤港澳大湾区等区域重大发展战略,聚焦智慧港口、智慧航道、智能船舶等行业转型方向,例如西江黄金水道提质升级、世界级港口集群建设。场景承接宏观战略与微观企业实践,具备行业生态关联性、标准引领性,核心解决跨部门跨区域数据打通、数据权属界定、行业标准适配不足问题,通过协同治理、统一标准、技术创新优化区域资源配置效率。

3. 微观场景: 可复制企业运营案例

微观场景直面企业运营痛点与用户真实需求,聚焦生产智能化、运营精细化、服务平台化,例如水上交通安全风险预警、船舶进出港智能调度、航运企业安全全流程管理。场景需求精准、迭代敏捷,可细化至单个业务环节、设备、人员行为,重点解决碎片化数据整合、个体数据精准采集、隐私合规使用问题;依托低代码、边缘计算快速落地,实现数据价值直接转化为企业运营效益。

(二) “需求—供给”双向联通协同机制

建立场景需求与数据供给常态化对接机制,实现需求侧动态牵引、供给侧能力支撑、双向平台匹配。

1. 需求侧: 分层动态需求识别

链主牵头开展全行业深度调研,分层梳理国家战略、区域产业、企业运营三类场景需求,从价值、覆盖范围、时效多维度定义需求指标,基于投入产出、技术可行性评估需求优先级,绘制行业高价值场景矩阵;搭建核心用户社区,通过AI问卷、线上研讨、场景测试建立用户全程参与的需求迭代机制,持续更新数据建设方向。

2. 供给侧: 全链条数据能力支撑

全面归集水运全产业链数据构建要素关联图谱,搭建行业可信数据空间,实现数据精准检索、标准化服务输出;将数据采集、标注、治理环节工业化、标准化拆分,依靠技术与制度双重机制保障数据合规流通;链主联合行业各方制定统一数据标准,针对战略公益、市场商用、个性化定制开发差异化数据集供给模式,推动行业主体共建共享、迭代更新数据集资源。

3. 双向协同: 供需对接生态平台

搭建统一供需对接平台,开放场景需求发布、数据集目录查询功能;完善政产学研用联合攻关协同机制,针对复杂场景开展联合研发;设计收益分成、风险补偿共享机制,构建利益共担、价值共创的可持续产业生态,实现数据供给与场景需求高效匹配,最大化释放数据要素乘数效应。

结语:

链主通过统筹资源、统一标准、开放治理能力、搭建协同生态,有效破解水运行业数据分散、质量低下、流通失序、场景落地困难等长期痛点,推动行业数据发展从粗放式资源囤积转向精益化价值运营。该链主模式不仅适用于水运行业,也能够为公路、铁路、航空等综合交通行业,以及制造、能源等其他产业的数据要素市场化配置提供可复制实践范式。

未来研究可进一步围绕链主主导下的数据要素收益分配机制、数据资产价值评估模型、跨行业数据流通协同规则展开深化,完善数据要素市场化配套制度体系,持续发挥高质量数据集赋能新质生产力、推动产业高质量发展的核心作用。

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 [N]. 人民日报, 2022-12-20(001). DOI:10.28655/n.cnki.nrmrb.2022.013348.
- [2] 惠宁, 史小蓉. 数据要素对企业新质生产力的影响效应研究 [J/OL]. 西北大学学报(哲学社会科学版), [2025-03-25]. <http://kns.cnki.net/kcms/detail/61.1011.C.20250313.1356.002.html>.
- [3] 袁天荣, 张佩雯, 刘瑞. 数据要素共享与新质生产力——基于有效市场和有为政府的机制分析 [J]. 经济与管理, 2025, 39(02): 1-11.
- [4] 肖闻. 数字新质生产力对商贸流通业绿色全要素生产率影响效应的实证检验 [J]. 商业经济研究, 2025(06): 10-14.
- [5] 郭莹. 链主企业数字化转型对制造业高水平“走出去”的影响——基于产业链国内关联与政府补贴的双重视角 [J]. 技术经济与管理研究, 2026(7): 110-117.

作者简介:

王惠(1980年—),女,汉族,辽宁大连人,高级工程师、工学博士、单位:大连东软信息学院、研究方向:大数据规划与管理、大模型研发。

白晶(1970年—),女,汉族,辽宁沈阳人,副研究员、工学硕士、单位:交通运输部水运科学研究院、研究方向:交通现代化管理、数据治理体系现代化。

董晓岩(1978年—),男,汉族,辽宁辽阳人,工程师、工学博士、单位:交通运输部水运科学研究院、研究方向:计算机科学与技术。

基金项目: 教育部人文社会科学规划项目(24JD710020)和大连市社科联基金项目(2024dlskzd248)。