文章编号：2097-5899(2025)07-0034-06　　中图分类号：U296　　文献标识码：A
DOI：10.16669/j.cnki.issn.2097-5899.202411220001
引用格式：马珺杰，王春雷，赵阳，等. 数据治理在重载铁路基础设施运维管理中的应用与实践[J]. 铁路物流，2025，43(7)：34-39，64.
MA Junjie，WANG Chunlei，ZHAO Yang，et al. Application and Practice of Data Governance in Operation and Maintenance Management of Heavy-Duty Railway Infrastructure[J]. Railway Logistics，2025，43(7)：34-39，64.

数据治理在重载铁路基础设施运维管理中的应用与实践

马珺杰[1]，王春雷[1]，赵　阳[2]，蔡俊峰[2]，杜　帅[2]，耿小亮[3]

(1. 陕西铁路物流集团有限公司 生产技术部，陕西 西安　710000；2. 陕西靖神铁路有限责任公司 生产技术部，陕西 榆林　719000；3. 北京全路通信信号研究设计院集团有限公司 运输指挥与信息技术研究院，北京　100070)

摘　要：在重载铁路信息化发展过程中，产生了海量的人工整理数据和线上系统数据，这些数据由于产生时间、产生系统、存储介质不同而导致数据质量存在问题，如数据口径不统一、不同系统指标值不一致等。数据质量问题不但无法正确表示当前业务现状，也给业务决策准确性带来影响。大数据治理技术从数据质量、元数据、主数据等方面对数据质量问题进行综合整治和监控，保障业务数据准确性，使数据准确、高效赋能运维管理。研究针对重载铁路基础设施运维数据治理，重点论述了数据标准化、元数据等多个关键治理体系的构建路径与方法。通过典型案例分析，展示了数据治理技术在设备告警诊断与数据共享2个业务场景中的实践应用，论证了该技术体系在保障数据质量完整性、强化业务决策支撑能力以及提升运营管理效能等方面的显著价值。

关键词：重载铁路；数据质量；数据治理；元数据；主数据；数据质量监控

在重载铁路信息化发展过程中，产生了海量的人工整理数据和线上系统数据，这些数据由于产生时间、产生系统、存储介质不同而导致数据质量存在问题。数据质量问题对业务影响巨大，不但会直接影响信息系统的运行效果[1]，而且无法帮助业务进行准确的运维决策[2]。数据治理技术分别从数据质量、元数据、主数据3个方面来解决数据质量问题[3]，经过数据治理，数据定义统一、取值准确，数据质量得到整体提升。高质量数据可以准确描述当前业务现状，反映业务问题，对重载铁路基础设施运维管理具有重大意义。

1　数据质量问题概述

1.1　数据质量问题分类

数据质量问题主要分为底层数据质量问题和使用数据质量问题2类。底层数据质量问题指数据本身数据标准、含义不统一[4]，直接影响数据是否可用，如数据命名不准确、取值范围错误、无度量单位、数据精度不准确等。使用数据质量问题指数据描述缺失或不准确，虽然不会对数据准确性造成影响，但是会在使用过程中产生错用现象，如业务描述缺失、技术描述缺失、管理描述缺失、无数据量级描述等。无论是哪种类型的数据质量问题，都会对应用带来负面影响，需要及时进行治理。

数据质量问题的产生首先是由于产生数据的系统较多，且系统之间相对独立。每个系统在设计之初很难系统梳理已有系统产生的各类数据，从而导致不同系统间的数据规范等存在偏差。其次，由于系统服务范围多为某业务流程，数据从定义到产出仅是使用该系统的用户较为了解，其他用户因无法得到数据全方位的描述，进而导致数据使用错误情况的发生。最后，

收稿日期：2024-11-22　修订日期：2025-04-16

由于缺少长期监控措施，未能及时发现数据错误并提醒修正，导致耗费较大人力物力治理好的数据再一次发生质量问题。

1.2　解决数据质量问题的挑战

数据质量问题产生的主要原因为数据产生和采集的方式不同、数据来源于不同的组织实体、数据存储结构和访问模式差异等[5]，解决数据质量问题的挑战如下。

（1）数据来源多且结构不同，如线下台账、电子台账、BIM系统、监测系统等，多源异构数据给数据质量治理带来了大量底层处理工作。

（2）数据规范不统一，如长度单位，A系统定义为km，而B系统定义为m。大量数据规范需要进行统一化处理。

（3）各类数据描述缺失，如数据所属业务、数据量级、数据产出频率等。描述缺失导致可能存在错用情况。

（4）缺少有效监测数据质量的方法，无法及时发现数据质量问题。

因此，大数据治理技术应使用先进治理方法论，利用数据中台技术和治理工具，解决数据质量问题带来的挑战，保障数据质量。

2　数据治理架构和主要模块

2.1　数据治理架构

数据治理架构主要分为规则层、治理层、监控层。规则层以数据为对象，在保证数据安全的基础上[6]，为不同数据制定相应的规则要求，并将这些规则储存在规则库中，指导数据开发和数据验证。治理层在规则库治理后的数据基础上，建立指标库、建设元数据和主数据体系，对主要指标进行统一管理，同时丰富数据描述。监控层采用监控脚本和用户反馈等方式，对数据库内的数据按照既定的规则库进行校验，并实现报警功能。数据治理架构如图1所示。

2.2　数据治理模块

数据治理模块主要包含数据规则库、指标库、元数据、主数据、动态质量检测、行为监测6个模块。

（1）数据规则库。数据规则主要目标是将一切可量化标准进行规范化处理，数据规则库存储这些数据规则。为不同字段制定的规则，如字段命名规则、缩写规则、实体编码、数据类型、数据长度、数据精度等。这些规则作为数据库上线前的评审依据，也可以作为上线后定期检测系统的规则输入。数据规则库是数据治理的基座，底层数据质量好坏，依赖于规则的准确性和全面性。数据规则多通过人工设置，经过不断积累，慢慢构建完成。

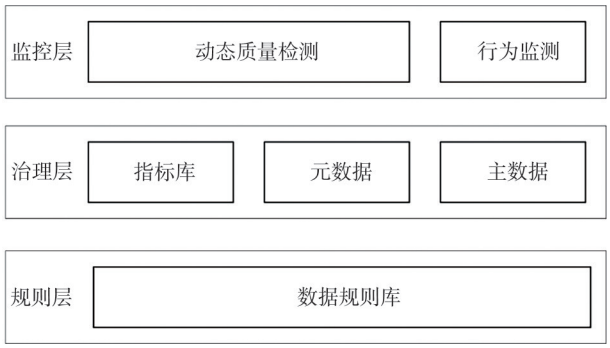


图1　数据治理架构

Fig.1　Data governance architecture

（2）指标库。指标库存储了各类表示业务状态、业务趋势和业务统计的指标，这些指标可为业务提供准确、实时的运营和运维情况[7]。指标分为原子指标、派生指标和复合指标3类。原子指标定义了最基本的指标，它无法继续拆解，任何指标都是基于原子指标产生。派生指标在原子指标基础上，扩展指标修饰词和时间周期，生成更为具象的指标。例如，原子指标是“员工数量”，当添加修饰词“男”、时间周期“2024年11月”后，该原子指标生成了派生指标“2024年11月男员工数量”。复合指标指的是多个原子指标或派生指标共同计算而得到的指标。例如，派生指标是“2024年11月员工数量”和“2024年11月员工总数量”，复合指标为“2024年11月男员工数量占比”。

（3）元数据。元数据是关于数据的数据[8]，它是对数据的补充描述，通过元数据，数据使用者可以更准确地了解数据的方方面面，更精准地判定数据是否符合需要，以及如何使用数据。元数据分为技术元数据、业务元数据和管理元数据3类。技术元数据客观描述了数据本身，如数据量级、数据产出时间、数据产出频率等；业务元数据从业务的角度对数据进行刻画，如业务含义；管理元数据从数据约束层面对数据进行描述，如数据负责人、安全等级、是否包含敏感数据等。

（4）主数据。主数据是对全量数据的一次提取，将业务重要的实体或者概念及其属性进行统一严格的定义，是最真实、最核心的数据资源[9]。主数据确保对业务有较大影响的数据正确和被统一使用，降低由于数据质量问题对业务带来的负面影响。例如公司数据中，公司名称、公司地址等都是非常重要的数据，在诸如签订合同等正式公文填写场景下，如数据出现错误，会给公司带来非常大的负面影响。

（5）动态质量检测。动态质量检测包含规则库、编排检测任务、统计质量结果、应急措施4个模块。规则库包含多条数据及元数据检测规则，检测规则可以是正则表达式、字段间比较、历史数据比较、统计数据比较等。编排检测任务通过一个或多个检测任务，使用规则库中规则对数据库内的数据进行定期或条件检测。统计质量结果在数据、表、时间等维度上对数据质量进行统计计算，得到数据健康度分数，来表示该数据质量的好坏。应急措施可配置告警和数据截断功能，对于质量较差的数据可通过报警方式通知其业务负责人和数据负责人，并自动分析问题位置；同时针对数据故障进行数据截断，避免错误数据传递到业务侧影响使用。

（6）行为监测。行为监测从使用层面对数据质量进行跟踪，通过开通数据反馈通道，收集用户使用过程中碰到的各类问题，分析定位质量数据范围和质量问题，并推动相关业务和数据负责人进行整改。

3　重载铁路基础设施运维数据治理

3.1　数据规则库建设

重载铁路基础设施运维数据覆盖了工务、电务、供电3个专业，以及设备、物资工器具、计划、作业等多个业务过程。这些专业和业务过程产生了多种不同结构和命名标准的数据。数据规则库建设首先对各系统进行摸排调研，整理归类不同数据的含义、类型、值域范围、数据长度、数据精度、枚举属性等信息，与业务侧共同完成这些信息的标准化工作，并将结果存储至数据规则库。在实现阶段，统一将不同结构的数据转化为结构化数据，并对导入数据进行数据清洗，对台账等进行数字化处理[10]，确保底层数据的准确性和可靠性。将数据规则库的规则应用到相应数据，检测是否满足规则，对不满足规则的数据进行纠正。数据规则示例如图2所示。

3.2　指标库建设

指标是一类特殊数据，它通常对重要的业务状态和业务趋势进行表达，对指标的规范化管理可有效提升这部分数据质量。指标库建设首先收集归类各采集系统及业务提出的指标集合，将这些指标进行拆解和重组，生成原子指标、派生指标和复合指标。数据开发工程师需根据这些指标定义进行指标开发。通过指标库建设，对指标的定义和计算逻辑进行规范定义，避免指标泛滥和错用。指标示例如表1所示。

3.3　元数据体系建设

元数据体系在数据治理和数据资产中扮演着重要角色，高质量的元数据为决策提供了必要的信息[11]。元数据体系建设范围包含数据表元数据及字段元数据。针对数据表元数据，通过采集数据表结构信息和后台实时监控表内容，收集数据表技术元数据，如中文名称、存储量大小、上游数据表等；通过走访调研及规章制度等方式，获取数据表管理元数据和业务元数据。针对字段元数据，由于其管理元数据和业务元数据与

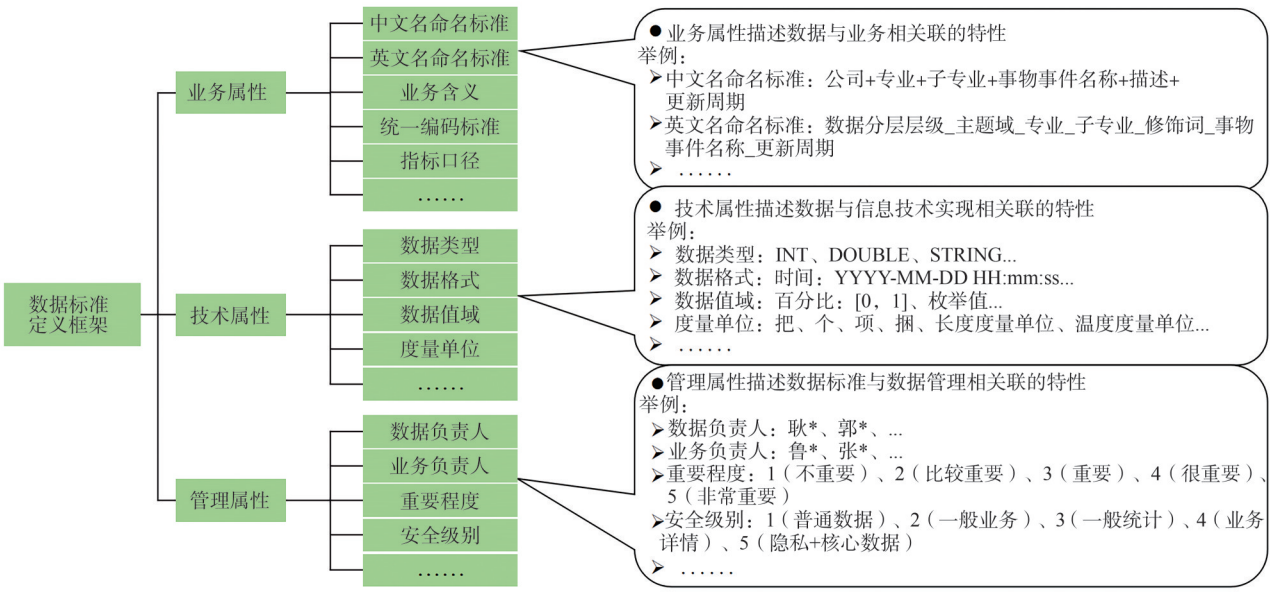


图2　数据规则示例

Fig.2　Example of data rules

表1　指标示例

Tab.1　Example of index

指标类型	英文名称	中文名称
原子指标	data_rpt_num	数据上报次数
	data_rpt_inr_dur	数据上报间隔
	prl_num	巡检次数
	prl_dur	巡检时长
	ct_num	检测次数
派生指标	data_rpt_inr_dur_avg	数据上报平均间隔时长
	prl_num_sum	巡检总次数
	prl_dur_sum	巡检总时长
	prl_dur_avg	巡检平均时长

数据表一致，故仅使用采集及监控程序收集字段的技术元数据。元数据类别如表2所示。

表2　元数据类别

Tab.2　Metadata category

范围	分类	元数据项
数据表	技术元数据	中文名称、英文名称、分区类型、使用频次分类、字段数、行数、存储量、存储类型、存储位置、生产频率、累计生产时间、最近生产时间、是否包含敏感数据、上游数据表
	业务元数据	业务定义、所属系统、所属专业
	管理元数据	业务负责人、数据负责人、安全等级
数据字段	技术元数据	中文名称、英文名称、字段类型、数据格式、长度、精度、值域、单位、状态、是否主键、是否唯一、是否可为空、上游字段

3.4　主数据体系建设

主数据提取数据中的重要实体或者概念，以及相应属性构建成主数据体系。在重载铁路基础设施运维中，基于工务、电务、供电数据，构建了组织架构主数据、基础主数据和设备主数据[12]。这些主数据及其属性经过严格的校对检查，并向整个数据体系和业务体系提供一份权威的数据内容。未来任何关于这类主数据的改动也需要经过严格的审批和审查过程。主数据示例如图3所示。

3.5　动态质量检测

动态质量检测根据数据特点和业务需求，制定数据动态规则，如日产出数据行数应大于阈值、工器具库存数不能为负、当日中文名称字段不合格比例低于20%等。同时，将动态规则和数据规则库中的静态规则合并，形成质量检测规则集合，集合中的规则由规则指标和规则阈值组成。

通过编写脚本代码和数据库查询语句等方式编写检测任务，获取规则集合中的规则指标，与规则阈值进行条件比较。多个检测任务可固定时间调度执行，

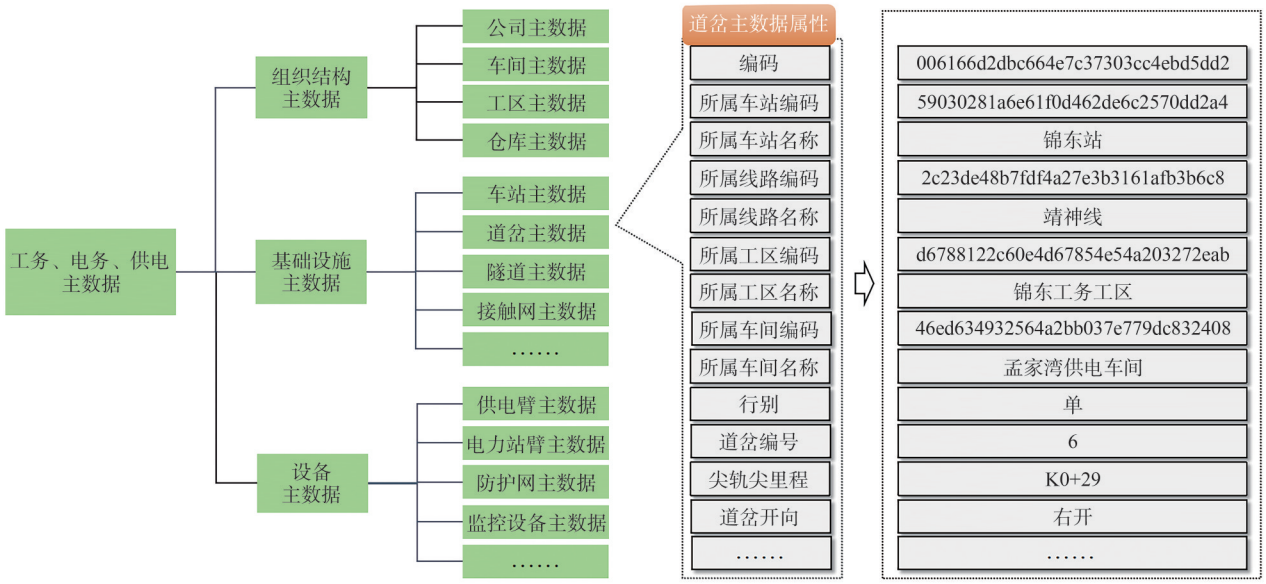


图3　主数据示例

Fig.3　Example of master data

如月度数据检查；也可以指定满足某个条件情况下动态执行，如在某个任务结束后执行。

检测任务执行完，对相应质量指标进行统计，重载铁路基础设施运维使用质量问题记录数和质量问题比例作为统计值，并设定指标阈值：质量问题记录数阈值100、质量问题比例10%。当这2个统计指标低于阈值时，可根据事先配置好的告警机制，通过短信方式通知到相应负责人自查解决。发生告警的数据也会停止参与业务数据计算和展示，一些影响不大的大数据量统计使用前一天数据替代，其他数据通知下游使用方出现数据问题，并在修复后通知恢复。

3.6　行为监测

由于数据治理后的数据还未完全应用在其他业务产品中，现阶段行为监测主要集中在邮箱问题反馈，未来会对行为监测做开放标准接口的研发，业务产品可调用行为监测，将用户在业务产品中关于数据的反馈同步至行为监测模块。

4　数据治理实践场景

4.1　基于元数据的告警诊断

重载铁路基础设施运维参与的部门较多，每个部门的工作范围和工作内容各不相同。当遇到某条告警数据时，传统做法是人为地基于经验去排查相关设备或上游设备状态，并且需要不同部门协作完成告警消除，其不同部门工作数据也无法有效关联。这种方式导致复杂诊断排查效率较低，诊断过程数据割裂，不利于告警经验的积累和使用。

数据治理通过对设备等进行统一编码[13]，将割裂的业务过程进行组合；同时，利用元数据体系中上游数据表和上游字段属性，建立整条链路的数据血缘关系。通过数据血缘关系，可以快速获悉某个字段是如何产生，进而可以形成回溯链路，展示链路节点中的数据内容，分析产生数据质量根因、数据质量影响范围，而不必基于经验去查看每个节点的数据内容，提升了排查问题效率。基于数据血缘关系，还可以获悉受问题节点影响的下游节点，并及时通知这些节点做好应急准备。

基于元数据的告警诊断，可以应用在工务、电务、供电等多专业的运维工作中。例如，电务通信专业在发现告警并进行排障过程中，需要人工分析告警原因和关联工区，然后通过电话沟通方式将排障工作转移到相应部门，该部门经过人为排查、解决问题后申请重新验证，值班人员验证无误后该告警标记解除。基于元数据的告警诊断，通过设备统一编码，打破部门间局限，将设备相关的上下游数据进行关联整合，建立起设备的数据链，链条上展示各个相关设备的实时数据，并可随时添加影响关系。当链条上某个节点产生告警，通信机房值班人员、车间工区员工都可以查看告警节点相关链条上其他节点的实时数据内容和诊断结果，通过数据发现设备异常，进而快速定位问题点，并进行特定排障，而故障排除后，设备实时状态也可以快速反馈到数据链。通过基于元数据的告警诊断，可极大提升告警诊断效率，提高问题定位准确性，在各部门和人员之间建立一致性数据共享能力，降低沟通成本，使告警诊断解决能力提升30%以上。

同时，基于数据治理数据血缘关系的告警诊断，可及时通知受影响部门做好预案，避免给业务带来更大的损失。未来随着信息化和人工智能的不断发展，问题从发现到排查再到下游通知，可以编排为标准业务处理流程，通过自动化的方式实现，使基础设施运维向着智能化方式迈进。

4.2　数据治理提升数据共享效率

由于数据来源多、数据结构多样化、数据标准不统一等原因，阻碍了数据在业务之间共享，主要表现在不知道数据在哪里、不知道数据能不能用、不知道数据怎么用等方面。

数据治理通过数据规则库对多源异构数据标准进行规范化，使相同类型的数据能够汇聚在一起，同时通过建设元数据和主数据体系，丰富了数据描述，对数据使用方的使用决策提供了数据支持。

数据地图便是基于上述内容构建的用于数据检索的数据应用产品。数据地图将元数据中的数据表名、字段名、业务定义等用于描述数据的信息组合并建立搜索索引，数据使用者可以通过搜索框进行数据搜索，也可以通过条件筛选完成数据获取，甚至可以精细到某项指标的检索，解决了“不知道数据在哪里”的问题。基于元数据和主数据，为数据表和字段建立身份名片，通过身份名片，数据使用者可以查看到数据量级、数据生产频率、数据所属业务部门等信息，通过这些信息数据使用者可以了解这些数据是否符合自身需求，解决了“不知道数据能不能用”的问题。基于管理元数据，数据使用者可以知道数据的安全等级，支持的获取方式等信息，解决了“不知道数据怎么用”的问题。

数据治理对复杂多样的数据进行了有效的规范化，并对这些数据补充了描述信息，通过可视化方式为数

据使用者搭建了简便好用的数据查询平台，提升了数据获取效率，同时减少了数据运维人员频繁解释数据使用者问题的场景，为业务系统提供了数据共享的便利，促进了数据的协同[14]。

5　结束语

数据治理通过数据规则库、元数据、主数据、数据质量检测等模块，改变了数据复杂不易使用的情况，建设了规范、准确、全面的数据体系和数据资产体系[15]，为数据在人工智能、业务产品、内部共享等方面提供了坚实的基础和便利性。通过数据治理体系，可以助力重载铁路基础设施运维解决大量问题，提升业务基础设施的运维水平。

参考文献：

[1] 蔡　莉，朱扬勇.从数据质量到数据产品质量[J].大数据，2022，8(3)：26-39.
CAI Li，ZHU Yangyong. From Data Quality to Data Products Quality[J].Big Data Research，2022，8(3)：26-39.

[2] 项姣姣.浅析数据治理在企业中的实施过程[J].互联网周刊，2024(9)：39-41.

[3] 吴信东，董丙冰，堵新政，等.数据治理技术[J].软件学报，2019，30(9)：2830-2856.
WU Xindong，DONG Bingbing，DU Xinzheng，et al.Data Governance Technology[J].Journal of Software，2019，30(9)：2830-2856.

[4] 王　喆.铁路大数据治理体系研究[J].网络安全与数据治理，2022，41(11)：30-35.
WANG Zhe.Research on Big Data Governance System of Railway[J].Cyber Security and Data Governance，2022，41(11)：30-35.

[5] 甘似禹，车品觉，杨天顺，等.大数据治理体系[J].计算机应用与软件，2018，35(6)：1-8，69.
GAN Siyu，CHE Pinjue，YANG Tianshun，et al. Big Data Governance System[J].Computer Application and Software，2018，35(6)：1-8，69.

[6] 董焕晴，何树坤，曹高辉.数字健康产业数据治理体系研究[J].现代情报，2024，44(9)：131-141，153.
DONG Huanqing，HE Shukun，CAO Gaohui. Research on Data Governance System of Digital Health Industry[J].Journal of Modern Information，2024，44(9)：131-141，153.

[7] 杨积成，邱浩楠，刘小鹏.供电企业全量指标库建设的探讨[J].百科论坛电子杂志，2024(10)：16-18.

[8] 黄学平，苏依婷，张永良.基于交通大数据的数据治理体系研究[J].通信与信息技术，2024(3)：115-119.
HUANG Xueping，SU Yiting，ZHANG Yongliang. Research on Data Governance System Based on Transportation Big Data[J].Communication & Information Technology，2024(3)：115-119.

[9] 戴　莉.浅谈企业数据治理[J].建筑设计管理，2023，40(8)：54-58.
DAI Li. An Insight into Enterprise Data Governance[J]. Architectural Design Management，2023，40(8)：54-58.

[10] 赵　颖，樊智伟，罗　浩，等.重载铁路通信设备数字化及全寿命周期管理技术研究[J].铁道运输与经济，2023，45(11)：48-57.
ZHAO Ying，FAN Zhiwei，LUO Hao，et al. Research on Digitalization and Lifecycle Management Technology of Heavy Haul Railway Communication Equipment[J]. Railway Transport and Economy，2023，45(11)：48-57.

[11] HAMZA E，MOHAMED R S，TAHER Z. Generic Model of Metadata Management System for Data Lakes[J]. International Journal of Metadata，Semantics and Ontologies，2023，16(4)：315-328.

[12] 杨连报，李　平，马小宁，等.铁路主数据全生命周期管理研究[C]//第十二届中国智能交通年会论文集.北京：电子工业出版社，2017：359-363.

[13] 单　伟，苏天祺，焦　晨，等.数据中台在重载铁路基础设施智能运维中的应用与实践探索[J].铁路物流，2025，43(3)：9-14.
SHAN Wei，SU Tianqi，JIAO Chen，et al.Application and Practice of Data Center in Intelligent Operation and Maintenance of Heavy Haul Railway Infrastructure[J].Railway Logistics，2025，43(3)：9-14.

[14] 徐　聪.大数据应用在云计算平台的优化部署与调度策略研究[D].北京：清华大学，2015.

[15] 钱　鑫，牛　庆，杜　平.数据治理视角下的数据资产管理体系构建[J].质量与认证，2024(8)：50-53.

（责任编辑：金　颖）

（下转第64页）