

领域知识插件化与多智能体协同的政务高质量数据集构建方法

禹芳¹, 杨琳^{2,3}

- 云赛智联股份有限公司, 上海 200233;
- 上海市大数据中心, 上海 200003;
- 华东师范大学数据科学与工程学院, 上海 200062

摘要

政务高质量数据集能够有效支撑政务服务、执法监管、城市治理等领域的智能化能力提升, 针对企业信用风险监管场景, 提出基于领域知识插件化与多智能体协同的高质量数据集构建方法。通过将领域本体和业务规则封装为可插拔知识插件, 实现架构与知识解耦, 支撑跨领域低代码迁移; 构建检测、规划、行动、评估的多智能体协作闭环, 实现数据质量管理全流程自治迭代; 建立可审计链式日志机制, 实现全链路血缘追溯与操作可追责。为政务领域高质量数据集建设提供可解释、可复用且具备跨域自适应能力的技术路径。

关键词

高质量数据集; 多智能体协同; 领域知识插件化; 政务数据质量; 知识图谱

中图分类号: TP399

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2026xxx

Domain Knowledge Pluginization and Multi-Agent Collaboration for High-Quality Government Dataset Construction

Yu Fang¹, Yang Lin^{2,3}

- INESA Intelligent Tech Inc., Shanghai 200233, China;
- Shanghai Municipal Big Data Center, Shanghai 200003, China;
- School of Data Science and Engineering, East China Normal University, Shanghai 200062, China

Abstract

High-quality government datasets can effectively empower intelligent capabilities in public services, law enforcement regulation, and urban governance. Focusing on the scenario of enterprise credit risk regulation, this paper proposes a high-quality dataset construction method based on pluggable domain knowledge and multi-agent collaboration. By encapsulating domain ontologies and business rules into pluggable knowledge plugins, this approach decouples the system architecture from domain knowledge, thereby facilitating low-code migration across domains. Furthermore, it constructs a multi-agent collaboration closed-loop of "detection, planning, action, and evaluation" to achieve fully autonomous and iterative workflows for data quality management. An auditable chained logging mechanism is also established to ensure end-to-end data lineage traceability and operational accountability. This method provides an

interpretable, reusable, and cross-domain adaptive technical pathway for building high-quality datasets in the government sector.

Key words

high-quality dataset, multi-agent collaboration, pluggable domain knowledge, government data quality, knowledge graph

0 引言

国务院《“十四五”数字经济发展规划》等文件明确提出，要提升公共数据质量，推动数据资源开发利用，加快数据要素市场化配置改革^[1]。市场监管领域同步推进企业信用风险分类管理改革，要求重点从企业基础属性信息、动态信息、监管信息、关联关系及社会评价等方面构建企业信用风险分类指标体系^[2]。然而，多源异构数据在汇聚融合中的质量瓶颈，已成为制约监管效能提升的关键堵点。

实践层面，构建政务高质量数据集面临三重困境。一是各部门数据质量参差不齐，标准不一，字段缺失与个性化用语等问题普遍存在，导致同一实体在不同系统中往往存在标识异构与语义歧义；二是数据共享仍面临困难，实体标识不统一、时空基准不一致使跨部门关联依赖大量人工进行数据对齐^[3]；三是治理手段智能化不足，传统规则治理难以处理语义歧义和上下文异常，且流程固化，缺乏自修复能力^[4]。这些困境折射出现有技术路线难以兼顾规则约束的确定性与语义理解的灵活性。以目前两种主流技术路线为例，传统ETL依赖规则脚本，能处理确定性要求，但面临规则组合爆炸且无法识别语义歧义的问题；大模型虽具备语义理解能力，却存在幻觉不可控、过程黑箱、成本高昂等缺陷。

针对上述问题，提出基于领域知识插件化与多智能体协同的政务高质量数据集构建方法。由大模型负责语义理解，规则引擎负责硬约束校验，插件封装领域本体与业务知识，智能体实现动态任务编排与自治闭环。主要贡献包括：1) 实现领域知识插件化，将领域本体、业务规则与别名库封装为可插拔的知识插件，通过注册中心动态加载与热更新，实现架构与知识解耦，支撑跨监管子领域低代码配置迁移；2) 设计多智能体协作闭环，构建检测、规划、行动、评估四类智能体协作闭环，分层扫描语义异常，动态调整任务优先级，在双重约束下递进执行并量化验收；3) 形成可审计链式日志机制，实现全链路血缘追溯与操作主体实名可追责。

1 相关研究

1.1 数据质量管理与数据集构建

在数据质量理论研究方面，Wang和Strong较早提出完整性、准确性、一致性、时效性等核心维度，形成了数据质量评估的经典范式^[5]。随着技术进步，质量提升手段从传统规则校验向智能化方法演进。Vale Silva等对公共部门数据质量研究进行系统综述，指出当前质量提升技术已发展到基于机器学习的异常检测与数据补全阶段^[6]。杨琳和朱扬勇提出了面向产品需求的全生产流程质量管控方法，为高质量数据产品的生产与流通提供了理论基

础^[7]。然而，上述研究多侧重于通用数据质量管理技术，对政务数据所特有的敏感性、权威性、强关联性以及复杂业务规则约束考虑不够充分，缺乏与具体监管业务流程深度融合的质量管理方案。

在政务高质量数据集建设实践方面，以企业信用监管为例，浙江省市场监管局运用机器学习构建信用风险模型，归集39个部门8.6亿条数据，从公共信用、治理体系、资本能力、质量能力、市场表现五个维度设计指标体系，验证了数据整合对信用监管的支撑价值^[8]。但其数据清洗环节仍依赖传统规则处理，未能有效解决企业名称语义歧义、跨系统状态矛盾及隐性关联缺失等深层质量问题。由此可见，现有政务数据在“聚数”与“用数”之间仍存在明显断层，亟需构建全过程精细化的数据集质量管理方法。

1.2 智能化数据集质量管理

近年来，大模型智能体（LLM Agent）技术为数据集质量管理的智能化提供了新范式。LAGA将图质量控制形式化为以数据为中心的流程，集成检测、规划、行动、评估四类智能体^[9]，实现了质量问题的自动化识别与修复；DataGovBench通过规划-执行-评估架构，为智能体在真实数据集构建 workflows 中建立了标准化测评基准^[10]；近年来，Yao等提出的ReAct框架通过协同推理与行动显著提升了LLM Agent的任务完成能力，其推理-行动协同思想为检测、规划、行动和评估类Agent架构提供了重要理论基础^[11]。此外，知识图谱增强的LLM推理研究表明，结构化知识可有效抑制模型幻觉并提升推理可解释性^{[12][13]}，但此类工作多聚焦于单轮问答场景，尚未与多智能体协

作的的多智能体质量管理流程深度融合。

上述研究主要存在三方面不足，首先，尚未针对政务数据集构建的特殊质量维度（如实体一致性、语义准确性、业务规则合规性）设计专用的智能体协作机制，现有通用架构难以直接适配监管业务的高精度要求。其次，缺乏将领域知识图谱作为“硬约束”嵌入智能体决策过程的理论框架，当前研究多将业务规则硬编码于智能体内部，导致跨场景迁移时需重写代码，可复用性较差。再次，智能体决策的可解释性与结果可审计性不足，大模型本身的幻觉问题与黑箱特性，使其在政务敏感数据场景中的应用受到制约，现有研究尚未在智能与规则间形成有效平衡机制。

2 政务高质量数据集构建方法

2.1 建设路径

本文提出“规则硬约束+插件化知识+智能体协同”的政务高质量数据集建设方法，建设路径见图1。包括数据采集、数据预处理、领域知识插件设计、多智能体协同治理和可信交付五大环节，本文重点阐述领域知识插件设计、多智能体协同治理和可信交付三个环节。

本文建设路径与传统ETL建设路径相比，核心差异主要体现在规则灵活性、语义处理能力、人工依赖度、跨域迁移成本及过程可解释性、质量闭环能力六个方面，见表1。传统ETL建设路径依赖硬编码规则，数据必须适配预设规则，规则维护成本高昂，发现质量问题后需人工返修，无自治闭环能力。本文方法采用插件化配置实现热加载更新，通过LLM语义理解与检索证据支撑处理歧义，形成人机协同的高风险节点人工审核机制，更换插件即可适

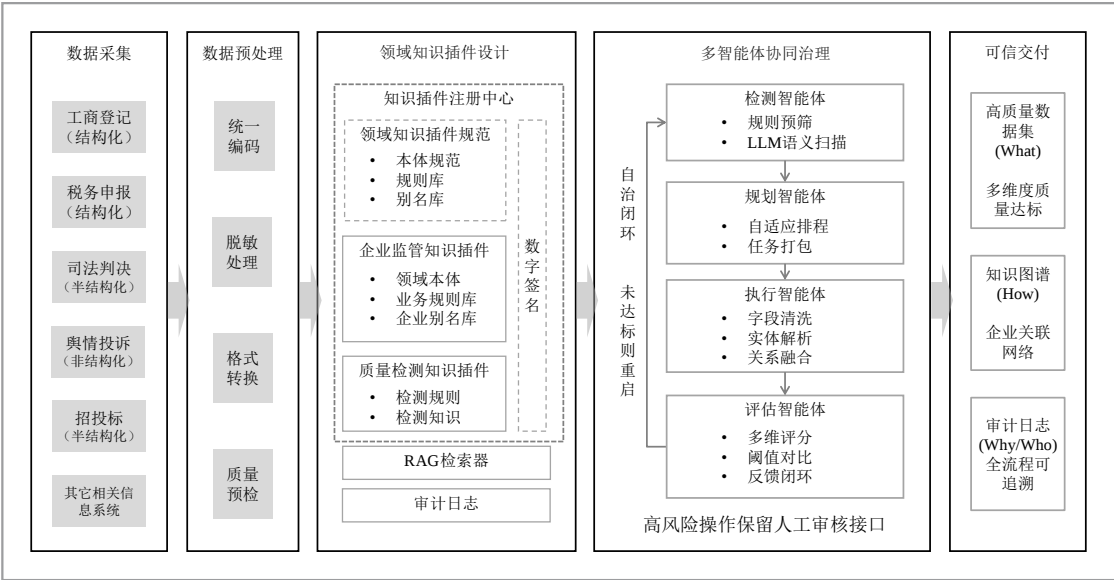


图1 高质量数据集建设路径

配新领域，并通过检索证据、规则校验和
审计日志三重可溯机制保障过程可解
释性。

表1 传统ETL路径与本文方法的核心差异

对比维度	传统ETL路径	本文方法
规则灵活性	硬编码规则,维护成本指数增长	插件化配置,热加载更新
语义处理能力	仅支持确定性转换,无法识别别名	LLM 语义理解+检索证据支撑
人工依赖度	质量问题发现后人工返修	人机协同,高风险节点人工审核
跨域迁移成本	需重写清洗脚本与映射规则	更换插件即可适配,核心架构复用
过程可解释性	规则日志可追溯	检索证据+规则校验+审计日志三重可溯
质量闭环能力	无自治闭环,依赖人工抽检	检测-规划-行动-评估自治迭代,关键节点人工确认

以企业信用风险监管为例，其本质上是一个“监测-预警-研判-处置-反馈”的闭环业务流^[2]。为支撑这一流程的智能化，本文从实体、关系、风险、知识四个维度进行需求建模。实体维度识别企业、人员、风险事件、资产等关键对象；关系维度标准化投资、担保、控制、供应、任职等核心关系；风险维度对风险事件进行分级分类，并结构化描述传导路径；知识

维度将《公司法》《企业信息公示暂行条例》等法律法规信用风险分类指标编码为插件本体中的推理规则与属性约束。基于上述需求，本文覆盖工商登记、税务申报、司法判决、行政处罚以及舆情投诉五大类核心数据源，定义关键实体与核心关系的数据抽取规则与关联路径，为后续多源异构数据的采集、清洗与融合提供业务基线。

2.2 领域知识插件设计

(1) 插件注册中心

本文将企业监管领域的专业知识封装为独立插件，智能体不硬编码任何业务规则，运行时通过插件注册中心动态加载^[15]。跨场景迁移时，同一套智能体架构仅更换插件即可适配新领域，实现架构与业务的深度解耦。

插件注册中心采用分层架构设计，见

图2。存储层管理插件包（含本体定义、规则库、别名库、版本元数据）及数字签名，确保插件来源可信；索引层维护插件标识符→版本→依赖关系的映射图，支持按领域标签（如“企业信用监管”“食品安全监管”）快速检索；调度层负责插件生命周期状态机管理，包含INIT（初始化）、LOAD（加载校验）、ACTIVE（运行中）、HOT_UPDATE（热更新）、SUSPEND（挂起）、UNLOAD（卸载）、ERROR（异常）七个状态。

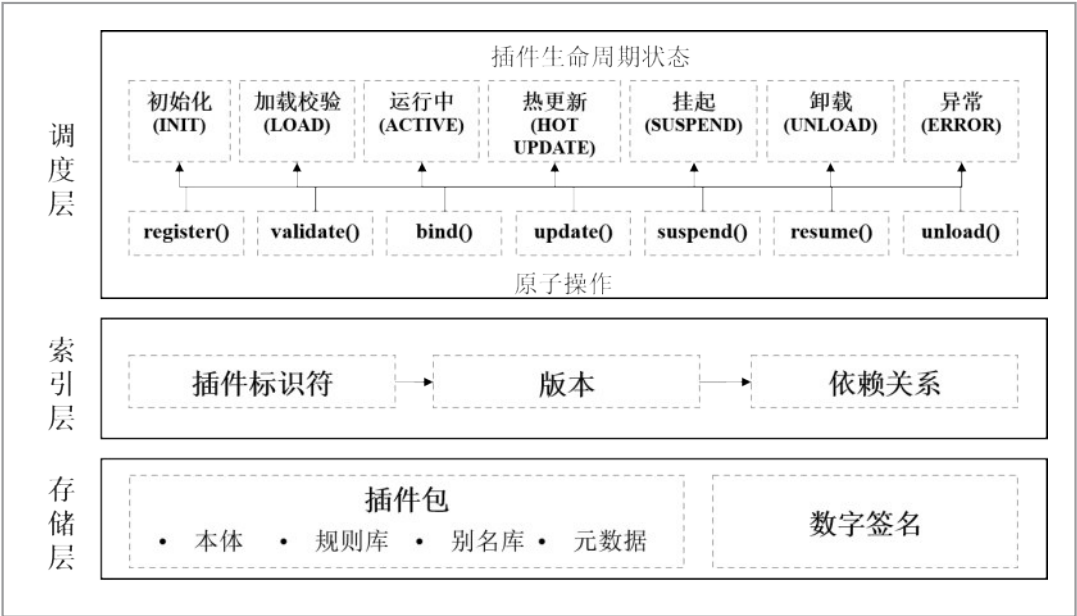


图2 插件注册中心

状态转移通过原子操作触发：register() 将插件注册至索引层；validate() 执行 LOAD 状态校验，包括接口签名一致性（方法参数类型、返回值类型匹配接口契约）、本体合法性（实体类型无循环继承，属性约束无矛盾，规则无死锁）、版本兼容性（新版本向后兼容旧版本数据格式）；bind() 将校验通过的插件绑定至智能体运行时上下文；update() 触发 HOT_

UPDATE 状态，在不中断智能体服务的前提下替换插件逻辑，通过双缓冲机制确保更新期间零停机；suspend() / resume() 控制插件临时停用与恢复；unload() 执行资源释放与日志归档。

热更新机制的核心是版本隔离与事务回滚。插件更新时，注册中心先加载新版本至隔离内存空间，执行冒烟测试（随机抽取100条历史记录验证输出一致性），测

试通过后原子切换引用指针，旧版本保留至垃圾回收周期。若更新后异常率超过阈值，自动触发rollback()回退至上一稳定版本，并记录异常堆栈至审计日志。该机制使业务规则变更无需重启智能体集群，支撑监管政策的敏捷响应。

(2) 核心接口规范

智能体通过标准化接口调用插件，实现架构与业务解耦。接口采用类OpenAPI的JSON Schema定义，核心接口规范如表2所示。

表2 插件核心接口规范				
接口	功能	输入参数	输出格式	调用方
getOntology()	获取本体定义的实体类型、属性约束、关系拓扑	domain_tag:领域标识符	JSON-LD格式的本体Schema	检测智能体、行动智能体组
ruleCheck(operation, context)	校验操作是否符合业务规则	operation:操作类型(清洗/对齐/融合); context:操作上下文(字段值、证据ID、置信度)	{valid: bool, violations: [], evidence_required: bool}	行动智能体组(修复后强制校验)
getAliasHints(rawName)	获取企业名称的候选别名列表	rawName:原始输入名称; similarity_threshold:相似度阈值(默认0.85)	{[alias: string, source: string, confidence: float]}	实体解析智能体
retrieveEvidence(query, filters)	从权威库检索证据	query:检索查询(实体名/关系/属性); filters:过滤条件(时间范围、数据源、置信度下限)	{[evidence_id: string, content: string, source: string, confidence: float]}	实体解析智能体、字段清洗智能体
getQualityThresholds()	获取质量阈值	dimension:维度标识(可选,空则返回全部)	{completeness: 0.99, accuracy: 0.97, consistency: 0.96, timeliness: 7, richness: threshold_obj}	评估智能体

接口设计遵循契约优先原则，所有插件必须实现上述接口定义的全部契约，返回值需符合JSON Schema校验，异常时返回标准化错误码（如PLUGIN_ONTOLOGY_INVALID、RULE_CONTRADICTION、EVIDENCE_NOT_FOUND）。该规范确保第三方开发的监管领域插件可无缝接入，形成生态化的领域知识复用。

(3) 领域本体设计

插件的核心是轻量级领域本体，定义实体类型、属性约束、关系拓扑及推理规则。与传统OWL全量本体工程不同，本文本体面向数据集构建任务定制，仅包含质量治理必需的规范。

本体核心定义示例如下：

实体类型集合

$$E = \{Enterprise, Person, RiskEvent, Asset, Industry\}$$

关系类型集合

$R=\{INVEST, GUARANTEE, CONTROL, SUPPLY, EMPLOY, LOCATE_IN, BELONG_TO\}$

属性约束集合

$A=\{name:String, uscc:String[18], reg_cap:Float\geq 0, paid_cap:Float\geq 0, status:Enum\{存续, 吊销, 注销, 迁入, 迁出\}, est_date:Date, industry_code:String[4]\}$

推理规则集合

$Rule=\{R1:paicap>Recap\rightarrow Violation(\text{“实缴资本不得大于注册资本”})R2:status=存续\wedge tax_status=非正常户\rightarrow Conflict(\text{“工商状态与税务状态矛盾”})R3:INVEST(A,B,r_1)\wedge INVEST(B,C,r_2)\wedge r_1>0.5\wedge r_2>0.5\rightarrow IMPLICIT_CONTROL(A,C)R4:CONTROL(A,B)\wedge CONTROL(B,C)\wedge isValidControlChain(A,B,C)\rightarrow CONTROL(A,C)\}$

(4) 质量指标定义

本文在 ISO/IEC 25012 数据质量模型^[16]基础上，结合政务数据强关联性特

征，面向企业信用风险监管提炼五维质量指标体系并形成质量检测插件，见表 3。为评估智能体提供量化验收标准。

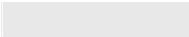
表3 五维质量指标体系

维度	定义	典型问题	治理目标
完整性	企业基础信息、年报数据、股东信息、司法记录字段齐全	统一社会信用代码缺失、股东信息漏报	关键字段非空率≥99%
准确性	企业名称、注册资本、经营状态与实际情况相符	注册资本数值错误、经营状态更新滞后	与权威源比对一致率≥97%
一致性	跨部门同一企业标识统一，状态无矛盾	工商“存续”vs税务“非正常户”	跨源一致率≥96%
时效性	变更记录、处罚信息更新满足监管周期	司法判决录入延迟、年报数据过期	关键字段更新延迟≤7天
关联丰富度	企业、股东、高管、关联企业、风险事件关系链完整	投资关系缺失、担保圈断裂、实控人隐匿	关系网络稠密可追踪

2.3 多智能体协同治理

多智能体协作遵循检测→规划→行动→评估的闭环流程，见图 3。检测智能体扫描数据异常并输出问题清单；规划智能体依据问题特征与历史治理效果动态生成

优先级权重，编排任务序列；行动智能体在插件知识约束下执行修复；评估智能体量化验收并触发反馈重启。四类智能体通过消息总线异步通信，共享插件注册中心与审计日志链。



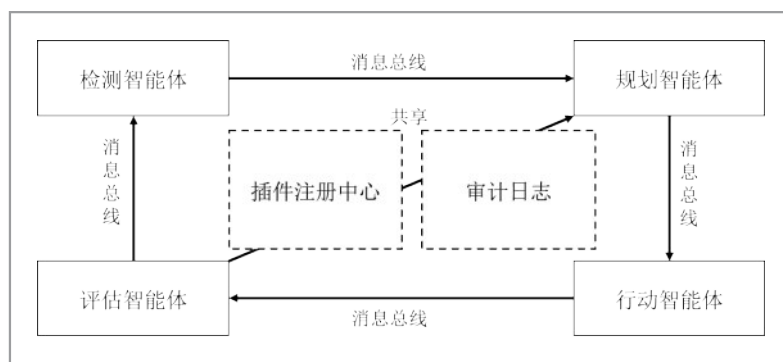


图3 多智能体协作流程图

(1) 检测智能体

首先由检测智能体负责分层扫描数据质量。第一层依据插件本体的属性约束进行规则预筛，如必填字段是否缺失，是否违反数值约束。第二层对非结构化文本调用大模型进行语义扫描，如识别企业名称别名歧义、经营范围语义偏差、跨系统状态矛盾。例如，针对原始记录：`{企业名称: 某某科技, 税务状况: 非正常户, 工商状态: 存续, 注册资本: 1000万, 实缴资本: 1200万}`，检测智能体输出结构化问题清单：实缴资本1200万 > 注册资本1000万，违反数值约束规则；税务状态“非正常户” ≠ 工商状态“存续”，跨系统状态矛盾；“某某科技”可能是“某某科技（深圳）有限公司”的简称，识别到企业别名。

(2) 规划智能体

规划智能体接收检测智能体输出的问题清单，结合插件中定义的字段依赖关系与业务优先级，生成有序治理任务序列。其核心机制是基于质量反馈的自适应权重调整。

每个问题的优先级得分 P_i 由三因子加权求和：

$$P_i = w_1 \cdot S_i + w_2 \cdot K_i + w_3 \cdot G_i$$

其中， S_i 为严重等级分（致命错误=1.0，语义歧义=0.6，格式偏差=0.3）反映问题的致命程度； K_i 为字段关键度（关键标识字段=1.0，普通业务字段=0.5，扩展参考字段=0.2）； G_i 为关联增益预期（孤立节点修复增益按网络连通性提升比例归一化至[0,1]）。 w_1, w_2, w_3 为动态权重，根据当前问题队列的特征动态计算，且满足 $w_1 + w_2 + w_3 = 1$ ，每项权重不低于最低保障阈值0.15。

每轮迭代结束后，评估智能体通过比较本轮与上轮的质量增益与耗时，形成自适应调节路径，持续优化策略。若单位时间完整性增益下降，则降低字段关键度权重，避免局部最优；若关联丰富度提升缓慢，则提升关联增益预期权重，优先处理孤立节点与断裂关系链；若致命错误残留率居高不下，则提升严重等级分权重，确保硬规则冲突不遗漏。这种“执行-评估-调整”的循环使排程策略随数据特征动态演化。

(3) 行动智能体

行动智能体由字段清洗、实体解析、关系融合三个子智能体组成，所有操作均受 ruleCheck 机制约束。

字段清洗智能体负责单字段级别的数

据修复，处理字段值对不对的问题，不涉及跨记录或跨来源的关联判断。修复策略分层执行，第一层为规则引擎直接修复，针对插件本体中已定义明确约束的字段（如数值范围、枚举值、正则格式），由确定性规则完成修复，确保过程可解释、结果可复现。第二层为大模型语义修复，仅当规则无法覆盖时触发，针对经营范围分类、地址标准化等需要语义理解的场景，将原始文本输入大模型，要求在检索证据支持下给出标准化结果，以RAG机制约束输出边界^[17]。

实体解析智能体负责跨来源同一实体的识别与对齐，处理不同来源的两条记录是否为同一个实体的问题。采用“检索-生成-校验”三阶段机制：检索阶段以企业名称、注册地址、法定代表人、统一社会信用代码为特征向量，从权威库召回候选证据，计算多字段综合相似度，缩小判断范围；生成阶段将原始名称与检索证据一并输入大模型，要求在有检索证据支持的前提下判断是否为同一实体，输出必须引用具体证据来源，杜绝无依据的语义推断；校验阶段设置硬约束，如大模型判断为“同一企业”时，必须满足以下至少一项条件：1）企业名称相似度超过阈值且注册地址行政区划代码一致；2）法定代表人一致且成立日期偏差在合理范围内，防止仅凭名称相似导致的误判。确认同一实体后，利用Jaccard相似度、编辑距离等方法检测并合并重复的事件报告^[18]，解决“同一处罚在多系统重复记录”的冗余问题。

关系融合智能体负责知识图谱构建与隐性关联发现，处理实体之间关系梳理、关系是否完整的问题，在字段已清洗、实体已对齐的基础上开展工作，不直接处理原始脏数据。显式关系映射阶段将清洗后

的结构化数据映射到知识图谱，通过插件本体获取关系类型定义，完成数据层与模式层的对齐。业务规则驱动的质量增强阶段利用图谱预置的业务规则自动发现数据矛盾，如“吊销企业不得有新增投资关系”，若存在近期投资记录则触发质量告警并提示人工核查，进行主动质量管控。隐性关联预测阶段采用RotatE模型学习实体与关系在复数向量空间中的旋转映射^[19]，预测缺失关系。

ruleCheck 机制是行动智能体的统一约束入口，其本质是插件层暴露的规则校验接口。行动智能体完成任何修复、对齐或融合操作后，调用ruleCheck将修复前后的字段值、操作类型及证据ID传入插件；插件依据本体中的属性约束与推理规则，对结果进行语法、语义及逻辑一致性检查。若触发规则冲突，或证据置信度低于阈值，ruleCheck返回校验失败信号，行动智能体必须回退该操作并将记录转入人工审核队列。该机制形成“智能修复-规则校验-异常回退”的约束闭环，防止大模型幻觉或规则遗漏导致的错误数据下渗。

(4) 评估智能体

评估智能体负责质量验收与迭代调控，不直接参与数据修复，而是对行动智能体的输出进行独立、客观的量化评判。评估智能体从插件接口获取五维质量阈值，对当前数据集进行全量扫描与量化评分。五维质量全部达标则进入人工终审环节；任一维度未达标则生成结构化反馈信号返回规划智能体。反馈信号包含未达标维度标识、问题分布统计及策略调整建议。

为防止无限循环，评估智能体维护迭代计数器与质量增益衰减监测。若连续三轮迭代中质量增益低于预设阈值（如单轮

综合得分提升 $<2\%$), 则触发收敛停滞告警, 暂停自治闭环并转人工研判。评估智能体不参与检测、规划、行动任一环节的具体决策, 其评分标准完全由插件阈值定义, 不受本轮迭代策略偏向的影响。这种算评分离设计确保质量评估的客观性与闭环调控的可靠性, 是政务场景可审计性要求的关键支撑。

2.4 可信交付机制

本文方法输出数据、知识、过程三类标准化交付物, 形成三位一体的分层交付体系。

(1) 结构化数据表: 以关系型表或列式存储格式交付, Schema 严格遵循插件本体定义, 包括企业主表、股东关系表、风险事件表、跨系统一致性标记表等。每条记录附带五维质量评分向量, 下游系统可按置信度阈值过滤使用。

(2) 知识图谱: 以属性图模型存储于图数据库, 节点与边均携带来源证据 ID^[20]。图谱承载三层关联信息: 一是显式关系, 即直接从权威源提取的投资、担保、控制、任职等关系; 二是隐性关系, 是经图谱嵌入模型预测生成, 单独标记置信度与人工审核状态, 未确认前不作为监管决策的直接依据; 三是风险传导路径, 可支持沿图谱拓扑传播风险信号, 辅助研判关联影响范围。

(3) 审计日志: 根据 GB/T 34945-2017《信息技术 数据溯源描述模型》国家标准, 记录操作主体、时间戳、输入输出快照、依据规则/证据 ID, 支持全链路血缘追溯。输出形态包括结构化审计日志表、证据包(原始检索证据、大模型推理输出、规则校验结果)及数据血缘图, 完整呈现原始字段→处理过程→输出字段的转换

链路。

三类交付物互相协同。结构化数据表提供标准化事实数据, 支撑定量分析与模型训练; 知识图谱揭示隐性关联与风险传导路径, 支撑定性研判与关联监管; 审计日志保障数据可信度与责任可追溯, 支撑合规审计与争议溯源。这种分层交付设计既满足机器可读的数据消费需求, 又满足政务场景对可解释性、可审计性的合规要求, 形成可信可追责的闭环保障。

3 跨场景适配分析

为推演本文方法在不同监管子领域的适配能力, 保持多智能体协作架构与插件注册中心不变, 仅将企业信用监管领域插件替换为食品安全监管插件。两场景在实体量级、关系拓扑复杂度、规则数量级上具有同构性, 见表 4, 具备通过插件转换实现架构复用的基础条件。

RAG 检索器同步替换为对应领域权威库后, 实体解析智能体可复用语义理解与证据检索能力; 关系融合智能体的图谱嵌入模型可通过插件获取新的关系类型定义, 适配食品溯源网络的结构特征。但需说明, 该推演尚未经过真实数据验证, 食品安全场景的实际治理效果受数据质量、权威库覆盖度及业务规则完备性等多重因素影响, 属于待验证的工程假设。

4 结束语

本文提出了一种基于领域知识插件化与多智能体协同的政务高质量数据集构建方法。通过检测、规划、行动、评估四类智能体的协作机制, 实现了大模型语义理

表4 企业信用监管与食品安全监管插件配置对比

对比项	企业信用监管	食品安全监管	说明
核心实体	企业、人员、风险事件、资产	生产企业、产品批次、餐饮终端、检验机构	实体类型同量级替换
关键关系	投资、担保、控制、供应、任职	生产、供应、销售、抽检、许可、溯源	关系拓扑同构调整
本体规模	10实体/20关系/10规则	8实体/15关系/12规则	规则复杂度同量级
RAG检索库	企业注册历史库、工商登记库	食品生产许可库、抽检记录库、溯源数据库	权威库同构替换
完整性(目标)	≥99%	≥97%	差距-2%
一致性(目标)	≥96%	≥94%	差距-2%

解与规则引擎硬约束的混合推理，以及基于质量反馈的自适应任务调度；通过领域知识插件化设计，将本体规范、业务规则与别名库封装为可热替换组件，使核心架构在不改动代码的前提下具备跨监管子领域的适配潜力。

本文工作仍存在一定局限。当前有效性以架构推演与逻辑论证为主，尚未在真实政务数据环境中开展大规模生产验证；大模型推理成本对高频批量数据治理场景的经济可行性构成挑战；涉及敏感政务数据的本地部署大模型与插件安全沙箱的协同机制尚需进一步工程化验证。未来研究将围绕以下方向展开：一是探索面向增量数据接入的轻量级智能体调度策略，在保持现有自治闭环能力的前提下降低全量迭代的计算开销；二是研究插件本体与业务规则的自动化生成方法，减少跨场景迁移时的人工领域知识编码成本。

参考文献：

[1] 国务院. 国务院关于印发“十四五”数字经济发展规划的通知[EB/OL]. 国发〔2021〕29号, 2021. https://www.gov.cn/zhengce/202410/content_6978910.htm.
State Council. Circular of the State Council on Issuing the “14th Five-Year Plan” for the Development of the Digital

Economy[EB/OL]. Guo Fa [2021] No. 29, 2021. https://www.gov.cn/zhengce/202410/content_6978910.htm.
[2] 国家市场监督管理总局. 关于推进企业信用风险分类管理进一步提升监管效能的意见[EB/OL]. 国市监信发〔2022〕4号, 2022. https://www.gov.cn/zhengce/zhengceku/2022-02/14/content_5673425.htm.
State Administration for Market Regulation. Opinions on Promoting Enterprise Credit Risk Classification Management and Further Improving Regulatory Efficiency[EB/OL]. Guo Shi Jian Xin Fa [2022] No. 4, 2022. https://www.gov.cn/zhengce/zhengceku/2022-02/14/content_5673425.htm.
[3] 高翔, 胡亚博, 胡盈盈. 政府促进社会数据共享的治理机制研究——基于“浙医互认”的案例分
析[J]. 公共管理学报, 2025, 22(3): 77-88.
Gao X, Hu Y B, Hu Y Y. Government governance mechanism for promoting social data sharing: a case study of “Zhejiang Mutual Medical Recognition” [J]. Journal of Public Management, 2025, 22(3): 77-88.
[4] ILYAS I F, CHU X. Data cleaning[M]. New York: ACM Books, 2019.
[5] WANG R Y, STRONG D M. Beyond accuracy: what data quality means to data consumers[J]. Journal of Management Information Systems, 1996, 12(4): 5-33.
[6] VALE SILVA M I, GOMES LUNA R N, CAVALCANTE A K B, et al. Systematic

- mapping of data quality in the public sector: dimensions, metrics and practices [C]//Proceedings of the XXIV Simpósio Brasileiro de Qualidade de Software (SBQS 2025). Porto Alegre: SBC, 2025: 153–164.
- [7] 杨琳, 朱扬勇. 高质量数据集产品的形态和生产流程研究[J]. 大数据, 2026, 12(2): 64–74.
Yang L, Zhu Y Y. Research on the forms and production processes of high-quality dataset products[J]. Big Data Research, 2026, 12(2): 64–74.
- [8] 浙江省市场监管局. 市场监管领域企业信用风险预警场景[Z]. 2022. https://credit.zj.gov.cn/art/2022/2/14/art_1229636059_1828.html.
Zhejiang Provincial Administration for Market Regulation. Enterprise credit risk early warning scenarios in market regulation[Z]. 2022. https://credit.zj.gov.cn/art/2022/2/14/art_1229636059_1828.html.
- [9] ZHANG Z H, LI X K, ZUO Y L, et al. When LLM agents meet graph optimization: an automated data quality improvement approach[EB/OL]. arXiv: 2510.08952, 2025.
- [10] LIU Z Y, HAN Z Y, SONG Y F, et al. DataGovBench: benchmarking LLM agents for real-world data governance workflows[EB/OL]. arXiv: 2512.04416, 2025.
- [11] YAO S, ZHAO J, YU D, et al. ReAct: synergizing reasoning and acting in language models[C]//The Eleventh International Conference on Learning Representations (ICLR), 2023.
- [12] JIANG J, ZHOU K, ZHAO W X, et al. Think-on-Graph: deep and responsible reasoning of large language model with knowledge graph[EB/OL]. arXiv: 2310.01089, 2023.
- [13] PAN S, LUO L, WANG Y F, et al. Unifying large language models and knowledge graphs: a roadmap[J]. IEEE Transactions on Knowledge and Data Engineering, 2024.
- [14] BUNEMAN P, et al. Why and where: a characterization of data provenance[C]//ICDT 2001. Berlin: Springer, 2001: 316–330.
- [15] OSGi Alliance. OSGi core release 7 specification[S]. 2018.
- [16] ISO/IEC 25012:2008.
Software and systems engineering — software product quality requirements and evaluation (SQuaRE) — data quality model[S]. Geneva: ISO, 2008.
- [17] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Advances in Neural Information Processing Systems, 2020, 33: 9459–9474.
- [18] ELMAGARMID A K, IPEIROTIS P G, VERYKIOS V S. Duplicate record detection: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(1): 1–16.
- [19] SUN Z Q, DENG Z H, NIE J Y, et al. RotatE: knowledge graph embedding by relational rotation in complex space[C]//International Conference on Learning Representations (ICLR), 2019.
- [20] RODRIGUEZ M A, NEUBAUER P. Constructions from dots and lines[J]. Bulletin of the American Society for Information Science and Technology, 2010, 36(6): 35–41.



禹芳（1982-），女，云赛智联股份有限公司 仪电智慧城市设计研究院高级工程师 数据科学部总监，主要研究方向为数据治理、智能化应用、智慧城市。



杨琳（1979-），女，上海市大数据中心高级工程师，华东师范大学数据科学与工程学院博士生，主要研究方向为政务智能化、数据质量管理、数据治理。

收稿日期: 2026-04-30
通信作者: 杨琳, linyang@shanghai.gov.cn
基金项目: 国家自然科学基金资助项目 (No.62537001 和 No.62377012)
Foundation Items: The National Natural Science Foundation of China (No. 62537001 and No.62377012)