基于数据整合的产业链、创新链、人才链、资金链整合机制探索与实践

王俊喆　曾博群　赵刚　朱迪
北京大学 光华管理学院　北京　100871

摘　要：产业链、创新链、人才链与资金链数据分散异构，制约战略情报融合。本文提出基于领域特定语言（DSL）的模块化多链数据整合机制，通过“接入—清洗—转换—拼接—建模”五阶段流程，利用 DSL 配置引擎驱动流式处理自动化接入与标准化建模，并基于 Flink、Kafka、Doris 等平台构建可视化配置与灵活调度框架。仿真实验表明，相较传统 ETL 方法，执行时间平均节约 40% 以上，拼接精确率与召回率均超 93%，数据缺陷修复率显著提升，扩展性能接近线性增长，为数字政府、区域情报及科研管理平台提供了可复用技术范式。

关键词：数据整合；领域特定语言；数据建模；自动化处理

1 引言

在数字经济和智能社会加速演进的背景下，面向未来产业发展的国家战略高度依赖于对多源异构数据的高效整合与智能分析能力。特别是在产业链、创新链、人才链、资金链等关键环节中，数据已成为驱动创新、资源配置与战略决策的核心生产要素。然而，当前上述“四链”之间的数据呈现出高度分散、异构性强、标准不一、语义不一致等典型特征，严重制约了面向战略情报的深度融合和价值转化。虽然情报学界和信息系统领域在多源异构数据整合方法上取得了诸多进展，但大多数方法面向静态结构化数据，在应对跨领域、跨平台、跨时空维度的数据整合需求方面仍存在显著不足（刘则渊和赵宏，2023）[1]。尤其在政府和科研机构主导的数据平台建设中，不同链条的数据具有各自独立采集逻辑和管理规则，数据之间缺乏共享机制和语义对齐手段，使得系统性融合和情报支持机制建设面临挑战（张晓林，2021）。

为此，我们设计了一套基于 DSL 配置机制的“四链”数据整合模型，为面向战略情报支持的数字资源整合提供一种可行路径。理论层面，这种面向多源异构“四链”数据的可配置整合机制可以补足现有情报整合方法在自动化与场景适配方面的不足。实践方面，我们期望推动图书情报学科在大数据基础上的“数据—知识—情报—决策”转化逻辑的实践化探索，也为数字政府与战略性新兴产业政策智能分析提供支持。

2 相关研究综述

“多源异构数据整合”已经成为当前情报学的前沿议题之一，诸多学者围绕数据整合的语义对齐、结构映射、质量控制等问题进行了研究。黄如花（2022）指出，当前数据整合的核心挑战不在于技术实现，而在于数据源语义异构与上下文差异带来的知识转化困难。张建平（2021）在区域知识服务体系研究中强调，建立跨域、跨级联动的数据融合机制是支撑情报服务走向“智能化治理”的关键基础。国际上，Xu et al.（2020）在对战略情报系统的数据架构研究中提出，应构建基于任务驱动的数据联邦系统，提升多维数据快速整合与关联分析能力[2]。可见，情报学界已逐步认识到数据整合能力是情报系统智能化升级的根基，尤其是在支撑政策预判、产业分析、技术预测等方面，其重要性愈发凸显。

产业链、创新链、人才链、资金链（“四链”）作为国家治理和未来产业发展的关键维度，近年来逐渐成为数据治理与平台建设的重点对象。产业链与资金链方面，Zhai et al.（2022）提出通过知识图谱与规则引擎结合技术实现产业链上下游数据自动关联，显著提升了产业情报分析效率。创新链和人才链方面，冯建梅和李晨光（2023）采用语义本体构建“科研人员—项目—成果”间的数据关联网络，为区域科技决策提供支撑。Chen et al.（2021）则从技术演化路径识别出发，提出多维创新数据融合方法，实现了高校与企业专利数据、论文数据的关联建模。尽管“四链”数据整合的研究逐渐兴起，但大多数方案仍以单链为主、人工配置为重，

缺乏通用、自动化、跨链适配的整合机制，对面向政策制定和宏观情报服务的系统性整合支撑仍显不足。

DSL 作为针对特定应用场景设计的编程语言，在数据处理、流程编排、模型配置等方面展现出了良好适配性，近年来逐渐被引入大数据与数据中台建设中。国外方面，Hudak（1998）较早提出了模块化 DSL 在任务配置中的理论框架；Mernik et al.（2005）系统梳理了 DSL 设计与实现的全过程，认为其在降低复杂系统配置门槛、提升工程复用效率方面具备独特优势。国内方面，王赟和张宏（2021）将 DSL 应用于政务数据流程编排，实现了跨平台业务数据的快速整合与服务调用，表明其在“数据驱动型治理”中的潜力 [12]。然而，情报学领域中对 DSL 方法的系统性引入仍较为稀缺，尤其缺乏结合“多链”复杂业务逻辑的通用型 DSL 数据整合框架。如何将 DSL 机制与流式处理、数据质量控制、语义对齐技术结合，仍是有待深入的交叉研究问题。

综上，现有研究在以下方面尚存不足。大多数据整合研究偏重系统设计与流程开发，缺乏与情报学中“知识组织”“智能发现”“任务驱动”的深层结合；针对产业链、创新链、人才链、资金链的跨链结构性整合研究较少，尚未形成统一建模范式；现有整合多依赖人工配置 ETL 流程，缺乏可复用、可迭代的参数化配置模型。为此，本文尝试构建一套基于 DSL 的“四链”多源数据整合机制，在保持灵活性与高适配性的同时，实现对多源异构数据的标准化处理、模型驱动建模与自动化集成、落地和复用。

3 研究问题与假设

围绕多源“四链”数据整合的核心需求，本文提出了三项待回答的研究问题和相应假设。研究问题一关注 DSL 驱动机制的整合能力。在处理产业链、创新链、人才链与资金链四条异构数据时，DSL 机制能否实现高效的一体化整合？研究问题二聚焦性能与效果比较。与传统 ETL 方法相比，DSL 机制在整体整合效率与拼接准确性（Precision 与 Recall）方面是否具有显著优势？研究问题三考察系统的规模适应性。当数据量不断增大时，DSL 机制是否能够保持近线性增长的整合效率，体现出良好的横向可扩展性？针对上述问题，我们提出以下三个研究假设用 [3]。

假设 H1：在相同数据规模下，基于 DSL 的多链路整合机制在单位时间内完成的数据处理量显著高于传统 ETL 流程。

假设 H2：从匹配结果的精确率（Precision）与召回率（Recall）指标来看，DSL 机制整合后的数据质量优于依赖人工配置规则的方法。

假设 H3：随着整合数据量的增加，DSL 机制的运行时长增长趋势近似线性，明显优于传统方法的指数型增长表现。

4 研究设计

我们围绕多源异构“四链”数据的整合需求，构建了一个基于 DSL 的可配置数据整合机制，以语义约束的配置语言驱动数据接入、清洗、建模与融合操作，形成支持高频更新、结构异构、标准多变的智能数据整合平台。

配置机制。系统围绕数据整合任务的五个核心要素（数据源定义、字段映射、转换规则、清洗策略与建模模式）展开，配置语言采用面向任务的规则式结构，支持可视化配置与文本式组合，通过结构清晰、语义明确的配置语言，提升非专业人员在整合过程中的参与度，实现配置的复用与版本管理。

关键模块。整合任务由 Flink 流处理引擎在 DSL 驱动下自动解析并执行，处理流程包括五个关键模块：数据接入模块使用 Kafka 作为数据中转通道将数据流接入系统，采用 Hash 分区算法实现并行数据调度（Stonebraker et al., 2005）；数据清洗模块基于 DSL 规则选择数据清洗方式。缺失值填补采用均值方法，异常值检测使用标准差阈值法；字段转换与标准化模块支持时间格式转换、字段归一、主键重命名等标准化操作；多链拼接与一致性验证模块利用主键匹配算法完成异构链数据的拼接整合 [11]；数据建模与存储模块将整合后的数据使用 Doris 进行建表与建模，并采用 Range 分区方式提升查询效率。

5 模拟实验

为验证所构建的多链数据整合机制在实际应用中的性能表现，本文设计并实施了一系列模拟实验。首先建构了 4 类模拟数据集，为贴合“四链”业务，实验使用 Python 的 Faker 库结合分布函数生成结构异构数据集，并引入模拟缺陷（缺失值、异常值、噪声字段）增强真实性。数据集属性如表 1 所示：

表 1：模拟实验的“四链”数据集基本情况

链条类型	数据量（条）	字段数量	缺失率	噪声比例
产业链	1,000,000	20	5%	3%
创新链	500,000	15	2%	5%
人才链	200,000	10	3%	2%

链条类型	数据量（条）	字段数量	缺失率	噪声比例
资金链	300,000	12	4%	4%

其次性能测试聚焦四方面指标表现。一是整合效率，即 DSL 机制在不同数据规模下的执行时间表现；二是整合准确性，反映整合结果的匹配精度与完整性；三是数据质量提升，即缺失值与异常值的自动修复能力 [4]；四是可扩展性，反映随数据规模增长的系统运行性能变化趋势。

最后选用 Talend Open Studio+SQL 脚本手动配置 ETL 流程作为传统方法代表，与 DSL 机制在相同资源环境下运行实验。实验平台为 8 核 CPU 与 64GB 内存；环境配置为 Flink 集群 +Docker+Kafka；分别执行 ETL 与 DSL 任务 4 轮（不同数据量）取平均值作为实验结果；保证两组数据源、网络环境、运行时依赖一致。

整合效率。图 1 展示了不同数据规模下，DSL 机制与传统 ETL 方法的运行时间对比结果。结果显示，DSL 机制在所有规模下均表现出较高运行效率，平均节省时间超过 40%，这一结果印证了假设 H1：DSL 机制在整合效率方面显著优于传统方法 [5]。

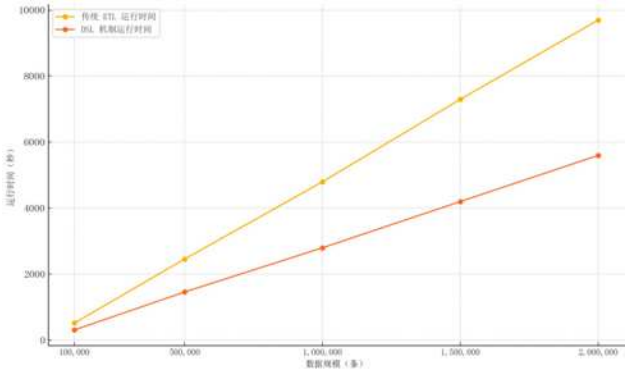

图 1：DSL 与传统方法运行时间对比图

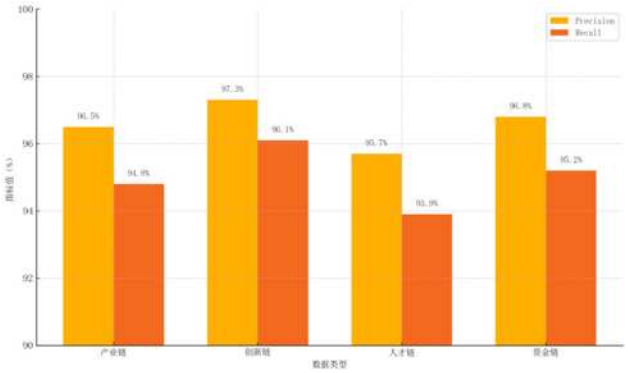

图 2：DSL 整合结果精确率与召回率对比

整合准确性。图 2 结果显示，DSL 机制通过统一字段标准化与键值映射策略，实现了更高质量的数据拼接效果，验证了假设 H2：DSL 机制整合后的数据准确性优于人工配置方法，降低了情报信息失真与误判风险 [10]。

数据质量提升。图 3 结果显示，通过 DSL 中可配置的清洗模块，系统可自动识别缺失字段、异常值分布并匹配适当修复策略（均值填补、标准差检测等）[6]。

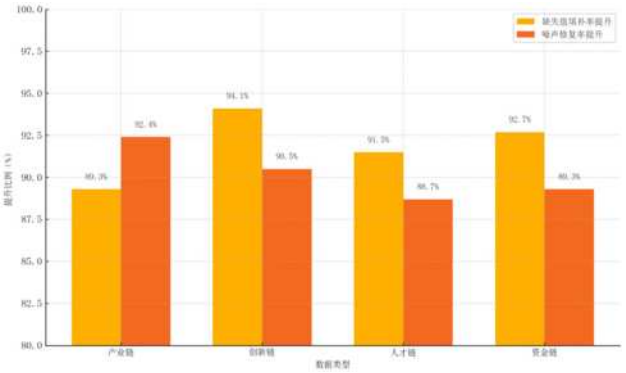

图 3：DSL 整合数据的质量提升效果

可扩展性。图 4 结果显示，DSL 机制整体呈现近线性增长趋势，而传统方法由于 ETL 流程缺乏并发机制与内存优化，在数据量翻倍时往往表现更像是指数级增长，验证了假设 H3：该机制具备良好的可扩展性，可适用于超大规模“四链”数据场景 [9]。

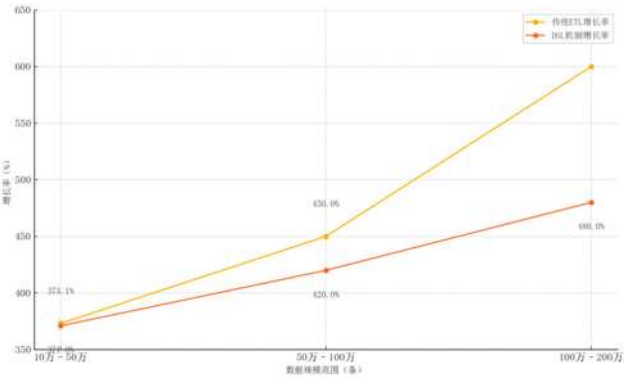

图 4：不同方法的扩展性对比趋势图

6 讨论与结论

本文构建了一个面向产业链、创新链、人才链与资金链（“四链”）的异构数据整合系统框架，并通过模拟实验从效率、准确性、数据质量与扩展性等方面进行了系统验证。结果显示，该机制设计在整合效率（平均节约 40% 以上）、准确性（精确率和召回率均超 93%）、质量提升与可扩展性方面均优于传统方法 [7]。理论上看，我们构建的数

据整合框架丰富了情报资源组织与管理的方法体系，通过明确将“产业链—创新链—人才链—资金链”四路数据作为情报源的核心构成，为学界关注新型情报源的整合机制提供了范式。实践上看，该整合机制可为政府决策分析、科研管理平台、产学研合作园区等多类机构快速构建“产业—创新—人才—资金”一体化数据平台提供技术路径。

本研究仍存在以下不足。首先，当前机制聚焦结构统一与流程自动，未深入处理语义异构、字段本体映射等复杂语义冲突问题；其次，如何在政务云、企业私域平台之间实现“可移植的整合方案”有待开展更广泛的平台适配研究；再次，数据权限控制、隐私保护、版本回滚等关键功能仍待加强[8]。最后，模拟实验在复杂政务、产业、区域情报应用场景中的实战表现仍需依托实际案例开展评估。后续研究可围绕“跨链协同—知识驱动—任务导向”三维展开，构建更贴近业务、具备智能语义感知能力的下一代数据整合框架，推动情报系统向智能化、可信化、自治化方向演进。

参考文献：

[1] 冯建梅 & 李晨光 . (2023). 科技人才链数据融合与可视化分析 . 现代情报 , 43(6), 36 – 42.

[2] 黄如花 . (2022). 异构数据整合中的语义冲突与解决机制研究 . 情报杂志 , 41(4), 19 – 24.

[3] 刘则渊 & 赵宏 . (2023). 基于语义融合的多源数据整合模型研究 . 情报理论与实践 , 46(4), 21 – 27.

[4] 王赟 & 张宏 . (2021). 面向政务数据流的 DSL 建模方法研究 . 电子政务 , (6), 88 – 94.

[5] 张建平 . (2021). 区域知识服务体系的数据整合机制研究 . 情报理论与实践 , 44(5), 67 – 74.

[6] 张晓林 . (2021). 数据驱动的国家知识治理与战略情报体系建设 . 图书馆论坛 , 41(5), 18 – 24.

[7]Chen, Y., Zhang, Q., & Wang, H. (2021). A data-driven method for innovation path discovery based on multi-source integration. Scientometrics, 126(4), 3421 – 3440.

[8]Hudak, P. (1998). Modular domain specific languages and tools. In Proceedings of the 5th International Conference on Software Reuse. IEEE Press.

[9]Mernik, M., Heering, J., & Sloane, A. M. (2005). When and how to develop domain-specific languages. ACM Computing Surveys, 37(4), 316 – 344.

[10]Stonebraker, M., Çetintemel, U., & Zdonik, S. (2005). The requirements of real-time stream processing. ACM SIGMOD Record, 34(4), 42 – 47.

[11]Xu, Y., Wang, J., & Liu, Z. (2020). Task-oriented data federation in strategic intelligence systems. Information Processing & Management, 57(6), 102384.

[12]Zhai, W., Wu, M., & Feng, J. (2022). Analysis of industrial product quality problems based on knowledge graph reasoning. In Proceedings of the International Conference on Computer Science and Communication Technology (ICCSCT 2022) (SPIE Vol. 12506, Article 2662045). SPIE.

作者简介：王俊喆，1988 年 11 月，男，四川省达州市人，汉族，讲师， 管理学博士，研究方向：大数据分析在组织的应用。

曾博群，1998 年 4 月，男，广东省广州市人，汉族，博士研究生，研究方向：数字经济 人工智能。

基金项目：科技部重点研发项目“面向未来产业生态的科技服务平台技术研发与应用”（项目编号：2022YFF0903200）。