# 数据产品的统一标识编码方法研究

吴俊[1]，谢文静[2,3]，熊赟[2,3]
1. 上海市城市数字化转型应用促进中心（上海市智慧城市建设促进中心），上海 200125；
2. 复旦大学计算与智能创新学院，上海 200438；
3. 上海市数据科学重点实验室，上海 200438

**摘要**
数字资源的高效管理与互联互通是实现数据赋能的基石。传统以元数据管理为核心的资源组织方式已难以适应跨部门、跨层级、跨业务的复杂协同需求，限制了数字资源的深度利用。数据产品是数据市场流通的一大类标的物，统一标识作为数据产品的基础，其研究与实践对于促进数据流通共享至关重要。在系统梳理标识编码方法研究进展的基础上，针对当前数据分析需求复杂化、数据多源异构化，以及数据智能体模式带来的多任务协同语义互操作等关键挑战，提出数据产品的统一标识编码方法。该方法通过引入机器可理解的高维语义特征，使数据产品成为智能体识别与操作数据的统一接口，为推动数字资源体系的标准化、智能化和协同化发展提供方法参考与标准化路径。

**关键词**
数据产品；统一标识；数据智能体；数据流通；数字资源

中图分类号：F49 文献标志码：A doi：10.11959/j. issn. 2096-0271.2026035

# *Research on unified identification for data products*

**Wu Jun[1], Xie Wenjing[2,3], Xiong Yun[2,3]**
1. Shanghai Application and Promotion Center of City Digital Transformation (Shanghai Smart City Promotion Center), Shanghai 200125, China
2. College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200438, China
3. Shanghai Key Laboratory of Data Science, Shanghai 200438, China

***Abstract***
The efficient management and interoperability of digital resources serve as the cornerstone for realizing data empowerment. Traditional resource organization methods, centered on metadata management, are increasingly unable to adapt to the complex collaborative demands across different departments, hierarchies, and business domains, thereby limiting the deep utilization of digital resources. Data products are key elements of circulation in the data market, and unified identification serves as their foundation; research and practice in this area are critical for facilitating data circulation and sharing. Based on a systematic review of the research progress in identification coding methods, this paper addresses key challenges such as the growing complexity of data analysis demands, the heterogeneity of multi-source data, and semantic interoperability required for multi-task coordination in emerging data agent models. It

proposes a unified identification coding method for data products. By introducing machine-understandable high-dimensional semantic features, this approach enables data products to serve as a unified interface for agents to recognize and operate data, offering methodological reference and a standardization pathway for advancing the standardized, intelligent, and collaborative development of digital resource systems.

***Key words***
data product, unified identification, data agent, data circulation, digital resource

## 0 引言

产品是市场流通的基石，是所有市场活动与价值交换得以发生的根本前提与载体。数据产品已经成为数据要素市场的高价值流通对象。不同于实物产品，经过加工的数据集、数据报告、数据服务、数据模型等均被作为数据流通的标的物，这些大小不一、模态多样的数据标的物辨识性差、流通受限。盒装数据作为数据产品的基本形态，为数据市场提供了可封装、可辨识的流通标的物[1-3]。如同图书拥有书号、电子设备具备序列号，产品标识是保障市场流通秩序的基础。长期以来，数据市场活动往往依据自身业务需求建立独立的数据编码体系，缺乏顶层设计的统一标识规范。这导致严重的数据异构性和语义割裂，制约了数据的流动与价值释放。因此，为数据产品赋予唯一性标识，是其在现代市场中有效流通与管理的前提。数据产品标识是一套涵盖资源定位、属性描述、权限控制与溯源管理等的综合性数字化基础设施，其内涵远超传统意义上的编码或编号。随着应用场景的深化，统一标识的需求从单纯的身份确认向语义表达与行为支撑转变。传统的标识仅需解决“这是什么”的问题，而面向未来的统一标识则需要回答“这与其他资源有什么关系”以及“可以对它做什么”的问题。这要求标识具备承载丰富语义信息的能力，能够支持机器可自动执行的语义推理与上下文关联。

当前，人工智能技术正加速从感知智能向认知智能演进，数据服务模式也从单向办理向多任务协同转变。在这一趋势下，“数据智能体”（data agent）作为一种能够自主感知环境、规划任务并执行操作的智能实体[4-7]逐渐受到关注，成为数据分析、利用和价值发现的重要方式之一。与传统的数据分析模式不同，数据智能体具备对复杂意图的理解能力和对异构数据的协同操作能力。面对数据分析任务，数据智能体通过理解用户的分析意图，自主规划分析流程，并调用相应的工具接口执行操作[8]，整个过程涉及对多个数据产品的查询、处理与分析。

因此，对数据产品内容的理解能力是数据智能体实现高效运行的关键。以盒装数据产品为例，其内容理解包括盒内数据与盒外包装两个层面：盒内数据类型多样，涵盖图像、音视频、文本、结构化数据等；盒外包装则包括产品登记证书，以及产品说明书、质量证书、合规证书等[2]。传统的标识方法难以提供关于盒装数据产品的数据内容（content）与模式（schema）的语义线索。为使数据智能体具有自主理解数据的能力，可借助向量嵌入（embedding）表示学习（representation learning）技术[9]，将数据产品的元数据、结构信息及核心内容映射为语义向量，进而构建融合语义特征的统一标识。该标

识不仅是数字资源的“身份证”，更是数据智能体理解数据内涵、建立资源关联、规划执行动作的认知锚点，有助于在数据分析任务中实现对数据含义的精准解析、对多源信息的语义融合，以及对分析流程的智能编排。

## 1 统一标识编码方法现状分析

传统标识编码方法主要聚焦于资源的定位与管理，主要解决数据的可达性问题，即数据能否被找到和访问；而新型数据标识编码方法面向数据智能体新趋势，致力于构建一个机器可自主认知、推理与协同的资源环境，以回应数据的“可理解性”与“可操作性”的核心挑战。本节将从技术体系演进、标准化进程及应用模式3个维度展开系统分析，以揭示标识方法如何从静态的资源管理工具，向支撑数据智能体生态演进的动态认知接口转变。

### 1.1 技术体系演进：从网络连通到机器理解

在信息化初期，资源标识的技术体系主要依赖互联网基础协议，如统一资源定位符（uniform resource locator，URL）和统一资源名称（uniform resource name，URN）。这一阶段的核心目标是实现资源的物理定位与网络连通，解决“数据在哪里”的问题。然而，这种基于位置的标识方法存在明显的局限性，一旦存储路径发生变更，链接便会失效，而且标识本身缺乏对资源内容的描述能力，计算机无法通过标识本身推断资源的类型或属性。

为解决上述问题，以数字对象标识符（digital object identifier，DOI）[10]、Handle系统[11]为代表的持久化标识技术被提出。这些技术通过引入解析系统，实现标识与物理地址的解耦，确保资源的长期可访问性。此后，随着语义计算与认知智能技术的兴起，技术体系向语义化标识演进。该阶段的标识不再是一串无意义的字符，而是通过结合资源描述框架（resource description framework，RDF）[12]与网络本体语言（web ontology language，OWL）[13]建立标识与本体库中的概念之间的映射。例如，当读取到一个标识为gov:enterprise/12345的数据资源时，基于本体定义，计算机不仅能定位到该企业的数据记录，还能推断出其属于“营利性组织”类别，并具备“纳税”“行政许可”等与之关联的属性。

然而，面对日益复杂的跨域分析需求，传统的语义标识难以支撑多源数据融合的复杂场景。与被动响应查询的传统系统不同，基于数据智能体的新型数据分析模式能够理解自然语言描述的业务指令，并自主拆解任务、规划步骤，最终协同访问多个数据产品完成分析。在此背景下，数据产品的统一标识编码体系正经历从“网络连通”向“机器理解”的转变。这一语境下的“机器理解”不再局限于基于预定义本体的简单推理，而是指数据智能体能够通过标识直接感知数据的高维语义特征（如结构含义、数据内容摘要），并基于这些具备认知能力的标识自动发现数据关联，通过任务规划和函数调用（function call）动态组合多个数据产品以完成复杂任务。这一转变标志着标识技术正从服务于人类检索的辅助工具，发展为支撑数据智能体自主作业的核心认知接口。

### 1.2 标准化进程：从语法统一到语义与行为协同

数字资源标准化最初侧重于语法的统

一。以政务场景为例，各国政府制定了大量的编码标准，规定了机构代码、行政区划代码、证照类型的编码规则。这些标准通过固定长度、特定字符集和校验位等方式，确保各系统间数据交换的基本格式一致性。例如，统一社会信用代码制度的实施，解决了法人单位身份识别的多头管理问题。然而，这种基于语法的刚性标准化在面对跨领域、跨部门的复杂业务协同及非结构化数据时，面临语义贫乏与状态割裂的挑战。以统一社会信用代码为例，虽然它能唯一标识一个企业主体，但无法直接反映该企业当前的经营状态或特定的行业属性。例如，对于“中小企业”的认定标准，工业和信息化部侧重于营收与人员规模，而统计部门更侧重于资产总额，这种定义上的差异导致语义歧义。本文提出的基于向量表示学习的标识编码方法，通过将不同的元数据定义与业务规则映射到同一高维向量空间，利用向量相似度计算实现语义对齐，消除跨部门协作中的理解偏差。同时，随着数据智能体的应用，行为协同标准开始成为研究的新焦点。行为协同标准是指规范数据智能体在访问、处理和交互数据资源时的动态行为模式与交互协议[14-15]。行为协同标准关注“智能体可以对数据做什么”以及“在什么条件下做”。在此阶段，统一标识不仅是资源的索引，更是承载数据使用协议（data usage policy）与访问控制策略（access control policy）的可计算化表达载体[16]。

### 1.3 应用模式：从数据静态发布到动态智能服务

早期的数据应用主要体现为基础数据库查询，标识的作用局限于索引键。例如，早期的政务数据应用主要体现为政府基础信息查询。随着“互联网+政务服务”的推进，应用模式转向了以事项为核心的流程驱动，标识开始用于串联不同部门、不同环节的业务数据，例如通过统一社会信用代码将工商注册、税务登记与银行开户等环节的数据打通。然而，在当前智能化治理阶段，应用模式正向数据智能体多任务协同的数据分析模式转变。智能服务场景下的需求往往是动态且非结构化的，这要求数据智能体不能仅停留在按预设流程调取数据，而必须能够自主理解对应的数据产品，并动态规划计算逻辑。在这种高阶应用模式下，传统的仅具备定位功能的标识无法满足数据智能体对数据内容理解与操作推理的需求。

Palantir 公司的本体论（ontology）思想为数据产品统一标识的演进提供了参考范式[17]。Palantir 的本体论不仅仅是数据目录的升级，它实际上构建了一个组织的数字孪生底座。该体系通过语义层（semantic）、动力层（kinetic）与动态层（dynamic）3层架构，将静态的数据记录、具体的业务行动与抽象的决策模型统一为一个可计算的业务操作系统。在这一架构下，数据不再是数据库中的“静态存储”，而是升级为能够被系统感知、被模型调用、被用户操作的“动态业务资产”。具体而言，借鉴 Palantir 模式，面向数据智能体的统一标识不仅应包含属性描述，更需绑定具体的操作方法。

## 2 数据产品统一标识编码方法

针对现有标识编码方法在语义缺失、动态适应性差以及难以支撑数据智能体协同等方面的不足，本文提出新型的数据产品统一标识编码方法，以适应基于数据智能体的数据分析趋势。该方法通过赋予标

识机器可理解的高维语义特征与逻辑关联能力，构建数据产品统一标识编码，使数据智能体能够基于上下文动态发现、评估并绑定最优数据产品，实现从被动响应需求到主动预判服务的转变。

## 2.1 标识编码语义表示方法

传统数据标识编码缺乏对数据内容的语义描述，导致数据智能体在处理异构数据时，难以跨越表名或字段定义不规范的障碍，多源联合查询的准确率因此受限。为此，标识编码语义表示方法构建起“多维特征抽取–向量化编码–语义空间映射”的生成机制，将数据产品的多维特征转化为机器可计算的稠密向量。

该机制的首要环节是多维特征抽取。不同于传统编码仅依赖元数据的模式，标识编码语义表示方法通过生成需要全面感知数据表的深层特征，以支撑模糊检索与推理。特征维度可以划分为3个层次：首先是元数据层（metadata level），涵盖名称、归属者及业务领域等描述性文本；其次是模式层（schema level），提取数据产品中的数据字段名称、字段类型及字段间的约束关系[18]；最后是内容层（content level），通过对数据产品中的数据记录进行采样，提取高频关键词或数据的统计分布特征。这种多维度的特征提取策略能够使标识直接反映数据的真实业务内涵，以应对表名命名不规范导致的分析失效问题。在此基础上，引入领域知识库作为特征构建的先验约束。利用领域知识库对提取的原始特征进行标准化清洗与语义增强，将经过领域知识库校准的规范化特征作为编码输入，提升后续生成的语义向量质量。

在特征抽取的基础上构建向量化索引。利用预训练语言模型，将上述多源异构特征映射为计算机可理解的数学表达。假设数据产品为 $T$，编码器 $E$ 将其映射为一个高维稠密向量 $\boldsymbol{V}_{\text{vec}}$，其形式化表示如下：

$$\boldsymbol{V}_{\text{vec}} = E\left(\text{Metadata} \oplus \text{Schema} \oplus \text{Content}_{\{\text{sample}\}}\right)$$

其中，$\oplus$ 表示特征的拼接或融合操作，生成的向量 $\boldsymbol{V}_{\text{vec}}$ 为该资源的隐式语义索引。隐式语义标识具备数学属性，即业务含义相似的资源在向量空间中的距离将趋于接近。这意味着，即便不同来源对同一类数据的命名完全不同，数据智能体仍能通过向量计算识别出它们在语义上的等价性或关联性。

## 2.2 语义标识与关联网络

仅拥有独立的语义标识尚不足以支撑复杂的跨域协同任务，还需要构建覆盖数字资源的数据关联网络，为数据智能体提供“定位–关联–推理”的全局导航能力，通过上下文关系发现价值。

### 2.2.1 资源层：以语义标识为核心的实体映射

资源层是关联网络的物理基础，其核心任务是将生成的隐式语义标识向量 $\boldsymbol{V}_{\text{vec}}$ 作为数据产品的标识映射。在此层级中，语义标识不仅是静态的索引键，更是动态属性的关联锚点。尽管初始建设主要聚焦于静态内容的语义关联，但随着持续迭代，可进一步扩展标识的属性维度，增加“调用频次”“数据鲜活度”及“安全密级”等动态标签。

### 2.2.2 关系层：多维语义网络的构建

关系层基于领域知识库与业务规则，

定义并构建连接不同标识数据产品的逻辑关联路径，形成资源关联网络。为了支撑数据智能体的复杂查询，设计5类核心关系边：一是“数据血缘”边，利用标识内容的相似度计算，连接具有派生关系的数据；二是“业务依赖”边，连接同一事项流程中先后产生的标识；三是“实体关联”边，通过schema模式对齐与语义匹配，将分散在不同库中的同一主体进行强关联；四是“权限传导”边，明确资源间的授权继承路径；五是“业务约束”边，将数据产品统一标识与相关的业务约束相连。通过关系层的连接，数据智能体可以沿任意关系边追溯上下游。

#### 2.2.3 推理层：面向智能体的自主逻辑推导

推理层构建推理引擎支持数据智能体基于关联网络进行复杂逻辑推导与隐性关联挖掘[19-20]。在该层级中，标识的向量特征被用于计算节点间的潜在关联概率，而预定义的业务规则库（如行政区划归属、产值排序逻辑）则用于约束推理边界。例如，在“产业链招商”场景中，数据智能体接收到指令后，首先将“寻找资方”意图转化为查询向量，在语义空间中匹配到“新能源央企”的候选实体；随后，推理引擎基于地理围栏规则自动推导出逻辑链条：“企业实体→(空间约束)→行政区划分→(业务匹配)→对应企业”。在此过程中，数据智能体依靠标识在向量空间中的距离判断数据的相关性。

综上所述，关联网络为数据智能体提供可计算、可导航的全域数据地图；数据产品统一标识为数据智能体识别数据内容提供指引，数据智能体依靠逻辑关联推演、规划数据操作。

### 2.3 标识即服务访问模式

基于构建的语义标识与关联网络，数据的治理范式从传统的静态目录管理升级为标识即服务（identity as a service，IDaaS）的动态访问模式。具体而言，当提交任务需求后，数据智能体首先解析其核心语义，将其映射至领域知识库中的概念实体并转化为查询向量；随后，通过计算查询向量与语义空间中标识向量的相似度，检索出语义关联的候选资源标识集合。这一过程中数据智能体无须遍历海量无关数据，即可快速锁定最优数据源。此外，针对意图不明确的模糊任务，支持基于关联网络的多轮对话交互，通过追问关键实体或约束条件补充上下文信息，进一步提升需求与资源匹配的精准度。

## 3 典型应用场景与案例分析

为系统验证本文提出的数据产品统一标识编码方法在真实场景中的适用性与推广价值，本节聚焦于典型领域，围绕数据产品的语义标识机制进行应用场景设计，并从案例分析角度展开讨论。

### 3.1 应用场景设计与讨论

从数据语义理解、多源关联分析与动态任务协同3个维度，对数据产品的统一标识编码方法的应用场景展开讨论。

首先，以政务数据共享开放场景为例。多源部门数据普遍存在语义异构问题，传统标识仅支持资源定位，数据智能体执行跨域查询时难以准确匹配数据表与字段。本文提出的数据产品统一标识编码方法将元数据、字段结构及内容摘要映射为高维语义向量并封装入标识，使数据智能体可

将自然语言查询映射至同一语义空间，通过语义相似度计算自动匹配最相关数据产品，实现跨系统互操作与精准检索，为政务数据开放与授权运营提供技术支撑。

其次，以区域经济分析场景为例。该场景需融合企业登记、税收、用电等多源数据以研判产业态势。传统标识缺乏语义关联表达，数据智能体难以自主发现可联合分析的数据产品。语义标识通过关联建模，发现数据间的主题相似性与属性互补性，使数据智能体可基于目标标识自动遍历候选集，通过邻近度计算发现关联资源并规划调用顺序，将孤立数据动态组合为分析链路，提升复杂任务自动化水平，推动数据共享向知识协同演进。

最后，在供应链异常诊断场景中，数据智能体需依次调用多接口并根据中间结果动态调整执行路径。传统模式依赖预定义规则，难以应对多变需求。本文提出的数据产品统一标识编码方法不仅承载了内容语义，还嵌入了函数调用接口描述以蕴含操作语义，使数据智能体可依据子任务需求筛选候选资源，并通过标识封装的接口参数与权限约束实现任务步骤与数据服务的动态绑定，提升对非结构化任务的适应性，为多智能体协同提供基础支撑。

上述3个场景系统呈现了本文提出的数据产品统一标识编码方法的多层次支撑作用。提出的标识兼具数据内涵载体与操作触发器双重功能。语义标识通过高维向量实现数据内容形式化表达，使数据智能体在统一语义空间中完成数据理解、关联发现与任务规划，为异构环境下的数据智能体协同提供可计算路径。该方法在工业互联网、金融风控等领域也有推广潜力，其核心在于将数据产品转化为数据智能体可认知、可操作的语义化接口，为数字资源体系的标准化与智能化建设提供方法参考。

### 3.2 案例分析

本文案例数据来源于真实的政务公示数据与商业公开信息，并据此构建了资金数据集与资产数据集。其中，资金数据集选取央企工商信息表，涵盖大型能源央企下属的三百余家子公司的名称、注册资本、所属省市县及经营范围。该数据集存在显著的行政区划层级缺失问题，部分记录仅标注市级或区县级名称而未标注省级行政区，且企业名称中往往隐含模糊的投资意向，传统基于关键词匹配的检索方法难以精准定位目标主体。资产数据集则选取昆山市光伏产业调研数据包，包含昆山市本地企业名录、行业产值统计表及专利信息表。该部分数据呈现高度的非标准化特征，表结构定义混乱，存在“纳税人识别号”同名列冲突、核心经济指标统计口径不一致以及无效占位符等情况。

针对行政区划标注不完整问题，数据智能体首先执行基于领域知识库的地理围栏过滤。该过程不依赖单一字段的匹配，而是对每条数据实体的全量字段（包括省、市、县及企业名称）进行扫描，并将扫描所得信息组合成包含省、市、县3级行政区划的完整地址描述。在此基础上，数据智能体利用内置的行政区划规则库对实体进行语义匹配。该机制不仅能够识别显式标注的省级行政区名称（如“江苏省”），还能依据知识库中的行政隶属关系，将仅标注为“区”或“市”但实际属于江苏省的企业实体准确归类。这一基于规则的过滤策略可有效剔除数据集中业务语义相似但地理位置不符的噪声数据。

标识内嵌企业经营范围、业务简介等语义特征。数据智能体调用编码器将分析查询需求（如“寻找负责光伏、风电投资的央企”）转化为高维查询向量，以计算其与候选数据产品的语义相似度，返回

“清洁能源”“电力开发”等与“光伏投资”等在语义上高度相关的数据。

在江苏省新能源投资央企的挖掘分析任务中，数据智能体能够识别包括华能（栖霞）光伏发电有限公司、华能江苏能源开发有限公司、华能国际电力江苏能源开发有限公司在内的核心投资主体。若仅采用传统的结构查询语言（SQL）关键词检索（如 LIKE '%江苏%'），将无法识别华能（栖霞）光伏发电有限公司与华能太仓发电有限责任公司，因为这两家企业的注册信息中仅标注了区县级地名，导致高价值招商线索的遗漏。若采用无约束的向量检索，虽能识别语义相关的光伏企业，但会将电投建能（嘉兴）新能源投资合伙企业（有限合伙）、定边国能新能源有限公司等外省企业错误纳入推荐列表，影响决策质量。

在江苏省昆山市光伏龙头企业的优选任务中，数据智能体生成了包含昆山康贝斯新能源科技有限公司、江苏中信博新能源科技股份有限公司、阿特斯阳光电力集团股份有限公司等企业的相关数据。值得注意的是，行业龙头企业阿特斯阳光电力集团股份有限公司在昆山市本地企业名录中的产值记录为无效数据，但在关联的行业企业数据中存有营收记录。数据智能体依托标识构建的关联网络，通过逻辑关联补全了该企业的经济指标，并基于产值硬指标优先、语义匹配度辅助的双视图评价策略，将其准确识别为产业链龙头企业。

## 4 总结与展望

构建科学有效的数据产品统一标识编码体系，是完善数据治理体系、构建数字空间秩序、促进数据市场流通的基础性课题。本文在数据智能体广泛应用的趋势背景下，提出具有语义表示的数据产品统一标识编码方法，为数据产品提供了具有可辨识性的统一标识，并通过案例分析进行了方法验证。在数据要素市场化配置的背景下，如何改进数据产品语义标识生成方法并形成系统化的统一标识编码体系，以更有效支持数据确权、流通交易与价值分配，同时进一步探索结合区块链与隐私计算的新型标识技术，成为数据市场机制探索的重要研究方向。

## 参考文献：

[1] 熊赟，朱扬勇. 面向数据自治开放的数据盒模型[J]. 大数据，2018，4(2)：21-30.
Xiong Y，Zhu Y Y. Data box：a novel data model for self-governing openness of data[J]. Big Data Research，2018，4(2)：21-30.

[2] 叶雅珍，朱扬勇. 盒装数据：一种基于数据盒的数据产品形态[J]. 大数据，2022，8(3)：15-25.
Ye Y Z，Zhu Y Y. BoxedData：a data product form based on databox[J]. Big Data Research，2022，8(3)：15-25.

[3] 熊赟，朱扬勇. 数据产品及其流通监管体系研究[J]. 大数据，2025，11(3)：98-107.
Xiong Y，Zhu Y Y. Research on data product and their circulation regulatory framework[J]. Big Data Research，2025，11(3)：98-107.

[4] Zhu Y Z，Wang L W，Yang C Y，et al. A survey of data agents：emerging paradigm or overstated hype? [PP]. arXiv preprint，2025，arXiv：2510.23587.

[5] Hong S R, Lin Y Z, Liu B, et al. Data interpreter: an LLM agent for data science[C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg: ACL Press, 2025: 19796-19821.

[6] Wang L, Ma C, Feng X Y, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345.

[7] Park J S, O'Brien J, Cai C J, et al. Generative agents: interactive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. New York: ACM Press, 2023: 1-22.

[8] Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: language models can teach themselves to use tools[C]//Advances in Neural Information Processing Systems (NeurIPS). New Orleans: Curran Associates, Inc., 2023: 42602-42618.

[9] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[10] Paskin N. Encyclopedia of Library and Information Sciences[M]. Oxford: Taylor & Francis, 2010: 1586-1592.

[11] Kahn R, Wilensky R. A framework for distributed digital object services[J]. International Journal on Digital Libraries, 2006, 6(2): 115-123.

[12] Pan S R, Luo L H, Wang Y F, et al. Unifying large language models and knowledge graphs: a roadmap[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3580-3599.

[13] Hogan A, Blomqvist E, Cochez M, et al. Knowledge graphs[C]//Proceedings of the ACM Computing Surveys. New York: ACM Press, 2021: 1-37.

[14] Wu Q, Bansal G, Zhang J, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation[PP]. arXiv preprint, 2023, arXiv: 2308.08155.

[15] Hong S, Zhuge M, Chen J, et al. MetaGPT: meta-programming for a multi-agent collaborative framework [PP]. arXiv preprint, 2023, arXiv: 2308.00352.

[16] Mavračić J. Policy cards: machine-readable runtime governance for autonomous AI agents[PP]. arXiv preprint, 2025, arXiv: 2510.24383.

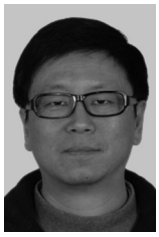
[17] Galis V, Karlsson B. A world of Palantir - ontological politics in the Danish police's POL-INTEL[J]. Information, Communication & Society, 2024, 27(13): 2438-2456.

[18] Wang Z L, Zhang H, Li C L, et al. Chain-of-table: evolving tables in the reasoning chain for table understanding [PP]. arXiv preprint, 2024, arXiv: 2401.04398.

[19] Zhu Y Q, Wang X H, Chen J, et al. LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities[J]. World Wide Web, 2024, 27(5): 58.

[20] Yao S Y, Yu D, Zhao J, et al. Tree of thoughts: Deliberate problem solving with large language models[PP]. arXiv preprint, 2023, arXiv: 2305.10601.

## 作者简介



**吴俊**（1970– ），男，上海市城市数字化转型应用促进中心（上海市智慧城市建设促进中心）高级工程师、主任，主要研究方向为政务业务数字化重构与流程优化等。

**谢文静**（2003– ），女，复旦大学计算与智能创新学院硕士生，主要研究方向为数据科学和数字经济。

**熊赟**（1980– ），女，博士，复旦大学计算与智能创新学院教授，上海市数据科学重点实验室副主任，主要研究方向为数据科学和数字经济。

**收稿日期:** 2026-02-25

**通信作者:** 谢文静,25213050422@m.fudan.edu.cn