基于可信数据空间架构的高质量数据集建设

张群洪[1]，李林[2,3,6]，杨堃[4]，陈爽[5]，张彬斌[4]，叶自燊[2,6]
1. 数字福州集团有限公司，福建 福州 350009；
2. 闽都创新实验室，福建 福州 350108；
3. 上海临港北京大学国际科技创新中心，上海 201306；
4. 福州数据集团有限公司，福建 福州 350207；
5. 福州市电子信息集团有限公司，福建 福州 350207；
6. 福建省数智双碳创新研究院，福建 福州 350014

摘要

高质量数据集是推动人工智能发展的关键基础，但现有建设模式存在质量评价偏静态预设、商业模式闭环困难、与实体经济脱节等现实难题。对此，提出并论证了一种高质量数据集建设的新范式：将高质量数据集建设深度融入数据要素市场，依托市场化流通与多维度用户反馈动态定义数据质量。该范式以“实体经济赋能为主、AI训练服务为辅”的数据要素双循环价值模型为核心，并依托“基于实数融合的可信数据空间”，通过全量流通图谱与数据流通大模型等，确保数据流通的可信、可控与可计量。在此基础上，设计了从数据资源入市到精准分发的全流程技术方案，并构建了可持续的商业模式与激励机制，为推动高质量数据集建设从“静态、供给驱动”转向“动态、市场驱动、生态化运营”提供了创新思路。

关键词

高质量数据集；可信数据空间；数据要素市场；动态质量评价；实数融合

中图分类号：TP315；TP391　　文献标志码：A　　doi：10.11959/j.issn.2096-0271.26084

High-quality dataset construction based on the trusted data space architecture

Zhang Qunhong[1], Li Lin[2,3,6], Yang Kun[4], Chen Shuang[5], Zhang Binbin[4], Ye Zishen[2,6]
1. Digital Fuzhou Group Co. , Ltd, Fuzhou 350009, China；
2. Mindu Innovation Laboratory, Fuzhou 350108, China；
3. International S&T Innovation Center at Lin-gang Special Area, Peking University, Shanghai 201306, China；
4. Fuzhou Data Group Co. , Ltd, Fuzhou 350207, China；
5. Fuzhou Electronics and Information Group Co. , Ltd, Fuzhou 350207, China；
6. Fujian Digitalization & Carbon-Neutrality Innovation Institute, Fuzhou 350014, China

Abstract

High-quality dataset is a critical foundation for advancing artificial intelligence(AI). However, significant practical challenges were identified in existing development models, including static and preset quality evaluations, difficulties in forming closed-loop business models, and a disconnect from the real economy. In response, a new paradigm for building high-quality datasets was proposed and justified: the development of high-quality dataset was deeply

integrated into the data factor market, and data quality was dynamically defined through market-based circulation and multi-dimensional user feedback. This paradigm was centered on“empowerment of the real economy as the main focus and AI training services as a supplement.” Trustworthiness, controllability, and measurability in data circulation were ensured under this paradigm using a “trusted data space based on real-virtual integration”, supported by full-flow circulation mapping and large models for data circulation. On this basis, a full-process technical solution was designed —from data resource market entry to precise distribution—and sustainable business models and incentive mechanisms were established. These efforts provided innovative ideas for shifting the development of high-quality dataset from a “static, supply-driven” approach to a “dynamic, market-driven, and ecosystem-based operational” model.

Key words
high-quality dataset, trusted data space, data factors market, dynamic quality evaluation, reality-date integration

0 引言

以生成式预训练大语言模型（Large Language Model，LLM）为代表的人工智能技术取得突破性进展，标志着人类正迈入前所未有的智能时代。大模型的性能边界日益取决于其训练所用数据的规模、质量与多样性。在此背景下，高质量数据集不再只是单纯的技术概念，更成为决定人工智能产业核心竞争力的战略性焦点，相关建设也成为全球科技竞争的重要前沿。

1 问题的提出与破题思路

1.1 相关概念与问题提出

2025年8月，在中国国际大数据产业博览会上发布的《高质量数据集建设指引》[1]指出，高质量数据集是“经过采集、加工等数据处理，可直接用于开发和训练人工智能模型，能有效提升模型性能的数据的集合”。其高质量主要体现在规模、安全等多方面，需静态与动态评价结合，且不同行业评价侧重不同[2]。分类上，可从数据模态（单模态、多模态）、模型阶段（预训练、微调、评估）、行业应用（通识、行业通识、行业专识）三个维度划分。

目前，有关高质量数据集的相关研究主要围绕概念特征、处理过程、评价指标等技术层面展开[3-7]，而有关商业模式与建设路径等深层次问题的剖析相对欠缺。现阶段，高质量数据集建设正面临深刻的系统性矛盾与商业可持续性危机。矛盾的根源在于“需求”与“供给”在逻辑与商业上的根本性错配。一方面，LLM训练对高质量数据的需求急剧膨胀；另一方面，支撑当前LLM发展的数据供给，其主体却源于一个近乎“免费”的体系。现有领先LLM的训练数据主要源于对互联网公开数据（如网页、开源代码、书籍文本与影音资料等）的大规模爬取与清洗。其成本集中于网络流量、治理开销及算力消耗，而非直接向数据内容本身进行市场化付费。这无形中塑造了一种产业惯性——高质量数据应近乎免费或低成本获取。

这种惯性与构建专有、深度、高价值数据集的商业化努力产生了直接冲突。当试图为金融时序数据、工业设备图谱、合规临床记录等专业领域建设高质量数据集时，其成本高昂（涉及隐私脱敏、专业标注、安全合规、权益清算等），但潜在采购方——AI模型研发机构——的付费意愿与

商业模式却建立在上一代“免费互联网数据”的认知之上。这使得许多高质量数据集项目陷入“投入巨大、回报渺茫”的困境，商业闭环难以形成[8]，且缺乏内生可持续动力。

与此同时，中国已从国家战略层面将数据定位为关键生产要素，其核心价值在于赋能实体经济转型升级。数据要素市场化配置改革的核心目标，是让数据在农业、工业、金融、医疗等实体产业中流通起来，解决真实业务问题，创造可衡量的经济效益[9-10]。这为高质量数据集的建设提供了一个更为根本和广阔的价值锚点：数据的“高质量”不仅可作为AI模型训练的“燃料”，更应体现在其对实体经济的赋能效能上。

因此，本文提出的核心问题是：如何打破“高质量数据集建设”与“不可持续的商业模式”之间的死循环？如何将高质量数据集的构建，从一个单纯为AI服务的、成本高昂的“项目”，转变为嵌入数据要素市场整体流通的、能够自我造血、自我演化的“生态过程”？这要求从根本上重构高质量数据集的建设范式、技术路径与价值实现逻辑。

1.2 本文研究思路

为解决上述问题，本文提出了一个根本性的思路转变：将高质量数据集的建设深度融入数据要素市场的整体运行框架之中，使其成为市场流通的自然结果而非预设前提。本文的核心论点是：数据的“高质量”不应由一套静态、抽象的学术或技术标准来预先定义，而应由其在真实市场流通与应用场景中产生的“用户反馈”与“效用结果”来动态定义与持续验证。具体而言，一个在工业故障预测中显著提升准确率的数据集，一个在供应链金融中有效降低风险的数据集，其“高质量”属性是由市场实效赋予的，具有客观的价值基础。

基于此，本文构建了数据要素“双循环价值实现”模型作为新范式的核心：

（1）主循环（实体经济赋能循环）：数据资源首先作为生产要素，在数据要素市场中通过合规机制配置到各类实体经济场景。在解决实际问题的过程中，数据的效用被验证、价值被实现，并积累丰富的应用反馈。这是数据价值实现的主渠道和商业闭环的基础。

（2）副循环（AI训练服务循环）：在主循环中经过充分验证、表现出高价值的数据，通过技术手段进行聚合、脱敏与产品化，形成面向LLM训练的专业数据集。此循环的价值来源于主循环已验证的经济价值溢出，和AI能力提升后对实体经济的反哺。

这一范式的实现，需要强大的底层基础设施作为支撑。本文以“基于实数融合的可信数据空间”[11]理论框架与技术体系为支撑。该体系通过“时空码”[12-13]支持实体对象和数据对象的可信身份锚定与事件追溯，通过“数联网”[14-15]实现跨域数据的标准化协议互联，通过“全量流通图谱”记录数据流通的全过程动态，最终借助“数据流通大模型”实现对数据价值和质量的智能评价。这一体系确保了数据在复杂、多主体市场环境中流通的可信、可控、可计量与可审计，为基于市场反馈的动态质量评价和可持续商业模式提供不可或缺的架构支撑。

1.3 论文结构安排

本文结构如下：第一节提出问题和破题思路；第二节剖析高质量数据集建设的

历史脉络、当前模式的局限及其认知误区；第三节阐述基于数据要素市场的新范式与双循环价值模型；第四节聚焦于实现该范式的关键技术路径与系统架构；第五节设计与之匹配的可持续商业模式与协同激励机制；第六节对新范式进行综合讨论，分析其优势、挑战与未来演进方向；第七节是结束语。

2 历史沿革与现状问题

2.1 演进脉络与路径依赖

高质量数据集的概念与实践并非随着LLM的诞生而出现，但其内涵与受关注程度却因LLM的发展发生了根本性转变。其演进可大致分为以下三个阶段。

第一阶段（2010年前）：任务驱动的封闭标注时代。这一时期的数据集建设主要服务于特定的、界定清晰的机器学习任务，如图像分类、自然语言理解等[16]。数据集的“高质量”体现为标注的精确性、一致性与任务相关性，其构建通常由学术机构或大型企业实验室主导，在小规模、封闭的环境下通过专家规则或众包标注完成。其价值主要在学术基准测试中体现，商业属性较弱。

第二阶段（2010—2020年）：互联网开放数据与预训练范式的兴起。随着深度学习，特别是自然语言处理中预训练模型（如BERT、GPT系列前期）的成功，训练数据的需求开始从“精准”向“海量”扩展。有研究者发现，模型性能的提升高度依赖于在大规模、多样化通用语料上的预训练[17]。这导致数据来源转向相对容易获取的互联网公开文本、图像和视频。这一阶段的“高质量”内涵开始偏向数据的规模、多样性与清洁度。与此同时，一个关键的商业模式特征在此确立：这些核心训练数据几乎不产生直接的版权或内容采购成本。数据获取的边际成本极低，主要成本集中于计算、存储与清洗的工程性投入。

第三阶段（2021年至今）：LLM时代的规模化与专业化矛盾。千亿级以上参数的LLM将第二阶段的数据需求推向极致，同时也暴露了其模式瓶颈。一方面，公开互联网数据的“低垂果实”已被采摘殆尽，数据质量开始制约模型性能的进一步提升。另一方面，要使大模型在金融、法律、医疗等专业领域真正可用，迫切需要深度、结构化、富含领域知识的高价值数据集。这恰恰与第二阶段形成的“低成本数据依赖”产生了直接而尖锐的矛盾。建设此类数据集需要解决隐私合规、权属清算、专家标注、多源融合等一系列复杂问题，成本高昂，但产业下游却缺乏成熟的付费模式来支撑这种成本。这种历史路径依赖与未来需求之间的断层，构成了当前高质量数据集建设困境的核心背景。

2.2 当前模式及其局限

当前，试图弥合上述断层的努力主要体现为3种模式，但这3种模式均存在结构性局限。

（1）科研驱动开源模式：由高校或研究机构主导，聚焦于创建新的学术基准。其优势在于推动前沿研究、定义评价标准，但也有其局限性：其一，数据规模通常有限，且场景高度简化，与真实商业环境差距甚远；其二，维护与更新滞后，难以持续反映动态变化；其三，缺乏商业化设计，无法形成价值闭环，高度依赖项目制经费，可持续性弱。

（2）企业内需驱动模式：大型科技公司或垂直行业巨头为自身业务或内部研发构建私有数据集。该模式贴近实际应用，数据质量与业务价值关联度高。其根本问题是封闭性：数据被视为核心商业资产，不进入外部市场流通，形成了“数据孤岛”。这不仅造成了重复建设和成本浪费，也使数据集本身难以通过更广泛的市场反馈实现检验与进化。

（3）专业外包服务模式：由专业的数据服务公司为企业客户提供定制化的数据采集、清洗、标注服务。该模式解决了特定场景下的短期需求，但本质上是人力密集型的“项目制”交付。其局限性在于：成本高昂，难以规模化；产出的数据集高度定制化，通用性和可复用性差；服务商与数据长期价值脱钩，通常只收取一次性服务费，缺乏持续优化数据质量的动力。

上述3种模式的共同局限在于，它们都将高质量数据集建设视为一个独立的、与数据要素市场主流流通割裂的“生产项目”。数据要么被封闭在学术或企业内部，要么作为一次性的外包成果交付。它们均未将高质量数据集建设置于一个持续运行、多方参与、价值可实时反馈与动态调整的市场化生态之中。因此，其所谓“高质量”往往是静态的、预设的，而非动态涌现的；其商业模式是线性的、一次性的，而非循环的、可持续的。

2.3 认知误区剖析

要想突破上述局限，必须先纠正3个根植于当前实践的认知误区。

误区一：目的误区——将“服务AI训练”视为首要甚至唯一目的。这一定位直接造成了商业模式的困境。如果高质量数据集的付费方只有预算有限且习惯于“免费数据”的LLM厂商，那么市场空间必然狭窄。即使将下游用户拓展到行业LLM范畴，其数量也是相当有限的。由此可见，数据的首要价值在于其作为生产要素，直接参与解决实体经济中的问题。工业参数数据用于优化能耗，信贷行为数据用于评估风险，其价值创造是直接且可货币化的。因此，服务于实体经济是数据价值实现的“主战场”，而服务于AI训练应视为价值溢出的“副产品”。

误区二：评价误区——迷信静态、通用的技术质量标准。目前评估数据质量的常见做法是制定一套涵盖完整性、准确性、一致性、时效性等维度的指标体系。这在封闭系统中具有一定意义，但在开放的市场环境中却可能失效。数据的价值是高度场景依赖和主体依赖的。同一组气象数据，对农业规划和航空调度的价值权重完全不同。因此，脱离具体应用场景和用户反馈的“绝对化高质量”实质是一种幻想。真正的评价权应交给市场，由数据在流通过程中产生的活性（交易频率）、效用（解决实际问题的效果）与口碑（使用者评价）等来综合决定。

误区三：建设误区——“先生产，后流通”的线性思维。传统模式遵循“采集—清洗—标注—封装—销售”的线性流水线。这种方式风险极高：投入巨大资源生产出的数据集，很可能因不符合市场需求而滞销。新的范式应倡导“在流通中建设，在反馈中进化”的敏捷模式。数据应以最小可行产品（minimum viable product，MVP）形态进入市场流通，通过真实应用获得反馈，并据此进行迭代优化。高质量数据集不是一个“完工”的固定产品，而是一个在市场中持续成长、动态演化的生命体，其建设过程与流通应用过程应融为一体。

3 新范式:动态质量评价与双循环价值模型

3.1 核心理念:动态质量评价

为破解高质量数据集建设的核心矛盾，本节提出一个根本性的范式跃迁：将数据质量的定义权与评价权，从预设的技术标准移交给持续运行的数据要素市场。数据的“高质量”不再是其固有的、静态的属性，而是其在市场流通与应用实践中动态形成的、被广泛认可的效用表征。这一理念包含以下3个层面的重构。

第一，评价主体从“专家”转向“市场多元主体”。数据的使用者（如工业企业、金融机构、医疗机构等）、加工使用者以及数据市场的运营方等都将通过其行为（包括交易、订阅、集成应用）与反馈（包含效果评价、需求迭代）参与数据质量的共同定义和评价。例如，一个在供应链金融活动中被多家机构频繁调用并有效降低坏账率的数据产品，其“高质量”标签就是由市场共同赋予的，具有客观的经济性证据和强有力的公信力。

第二，评价标准从“静态指标”转向“动态效用”。需要摒弃用一套通用技术指标评价所有场景的尝试，转而构建一个基于市场行为的动态评价函数。其核心输入变量应至少包括：(1) 流通活性(liquidity)：数据产品在指定周期内的访问量、订阅数、交易频率与金额，反映了市场的普遍关注度与需求强度。(2) 场景效用 (utility)：数据在具体业务场景中应用后，产生的可量化效能提升。例如，在生产优化中降低的百分比能耗，在精准营销中提升的点击转化率。(3) 网络效应(network effect)：该数据与其他数据产品组合、融合后，催生新解决方案或提升原有方案性能的能力，可体现数据的可组合性与生态价值。

第三，评价过程从“终点检定”转向“持续演进”。数据的质量分数不再是一个发布时的固定值，而是一个随时间、应用场景拓展和反馈积累而不断更新的动态值。市场成为数据质量的“实时熔炉”和“进化引擎”，持续淘汰低效数据，淬炼和凸显高价值数据。这也反过来证明了建立一套全局性可见的数据可信流通与评价体系的必要性[11]。

3.2 双循环价值实现模型

为建立可持续的商业模式，本小节提出一种“实体经济赋能+AI训练服务”的数据要素双循环价值模型，如**图1**所示。该模型旨在打破当前高质量数据集价值实现的单一路径依赖，构建起一个主副分明、有机协同的复合价值网络。

3.2.1 主循环:实体经济赋能

这是数据要素价值实现的根本通道，也是高质量数据集建设成本的核心承担者。其具体流程如下。

(1) 数据资源入市：原始数据经过必要的脱敏、标准化和产权登记后，作为数据资源进入数据要素市场。

(2) 市场配置与场景赋能：市场通过供需匹配机制，将数据资源配置到工业制造、现代农业、金融服务、医疗健康等实体经济的具体应用场景中。数据在这里作为关键的生产要素直接参与价值创造，例如，优化生产工艺、预测设备故障、评估信贷风险、辅助临床诊断等。

(3) 价值实现与反馈生成：数据应用产生直接或间接的经济收益（如成本节约、收入增长、风险降低）。同时，应用过程自

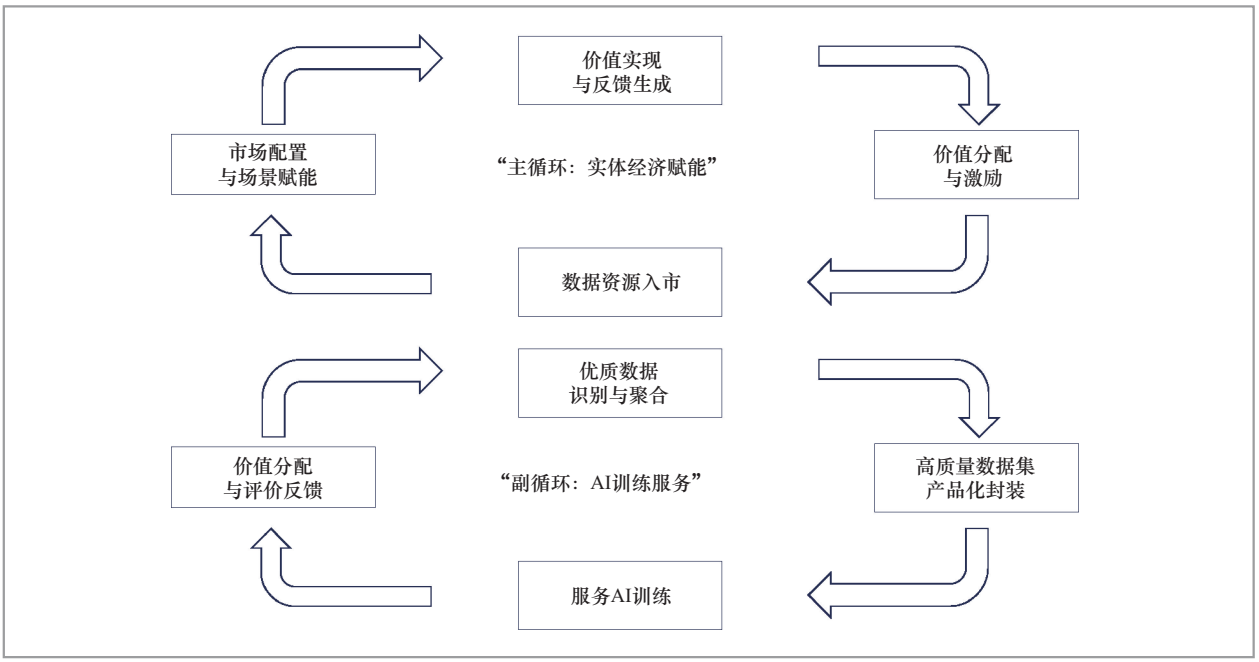

图1 数据要素双循环价值模型

动生成丰富的有效性反馈（如准确率、稳定性、业务指标提升度等），这些反馈被回传至数据要素市场，用于更新该数据的动态质量评价。

（4）价值分配与激励：基于合约与市场规则，应用数据产生的经济收益在数据提供方、数据加工方、平台运营方和应用方之间进行分配。数据因其产生的实效获得市场回报，形成商业模型闭环。经过市场验证、具有高效用评分的数据，其市场估值与交易价格随之提升。

3.2.2 副循环:AI训练服务

该循环不独立承担主要的成本回收任务，而是作为主循环价值溢出的承接者和增强器。其具体流程如下。

（1）优质数据识别与聚合：数据要素市场中的“数据流通大模型”自动识别在主循环中表现出高流通活性、高场景效用和高用户评价的数据（资源）。这些数据本质上是经过市场“淘洗”和“验证”的精华。

（2）高质量数据集产品化封装：将上述分散的高价值数据，在严格保障产权与隐私安全的前提下，由专业的数据服务商进行技术聚合、任务对齐与格式标准化，封装成适用于AI模型训练的高质量数据集产品。

（3）服务AI训练：由专门的数据服务商将这些数据集提供给AI模型研发机构，用于模型预训练、领域微调或专项能力提升。此时的数据集已非“原始原料”，而是承载了实体经济领域知识、经过市场效用背书的“精炼原料”，其价值基础坚实。使用高质量数据集训练后的AI模型，其增强的领域能力可以通过服务等形式再反馈到实体经济。

（4）价值分配与评价反馈：AI模型研发机构因赋能实体经济而获得相应的收入，其中的一部分再向高质量数据集产品生产、

流通各环节的相关参与方进行分配。同时，AI模型研发机构作为下游用户对上游产品（高质量数据集）进行质量、效用等综合性评价，评价结果将直接影响该数据集产品的市场估值和交易价格。

3.2.3 双循环协同关系

主循环是“主干”和“根基”，负责创造核心价值、承担主要成本、验证数据质量；副循环是“副产品”和“放大器”，专门针对AI模型进行二次转化与价值放大。二者的协同实现了数据价值的最大化——数据首先在实体经济中“建功立业”，证明自身价值；然后将其“成功经验”（体现为高质量数据集）赋能给AI模型，AI模型能力提升后进一步服务实体经济。这从根本上解决了高质量数据集建设的商业模式问题——其成本由广阔而刚性的实体经济数字化需求支撑，而非单纯依赖于AI模型研发的有限预算。

3.3 数据要素市场解决方案

上述动态评价与双循环模型的顺畅运转，依赖于一个高效运转的数据要素市场，其具体实现形态就是可信数据空间。根据国家数据局发布的相关政策文件，可信数据空间是“基于共识规则，联接多方主体，实现数据资源共享共用的一种数据流通利用基础设施；是数据要素价值共创的应用生态；是支撑构建全国一体化数据市场的重要载体”[18]。

李林等针对可信数据空间进行了较为系统的研究[9-12,16]，率先提出“基于实数融合的可信数据空间”整体解决方案，为有效破解数据要素规模化流通这一全球性难题提供了创新思路。该解决方案的主要内涵概括如下。

3.3.1 制度规则设计

根据数据要素的基础特性，构建“数据资源持有权（对应数据母本）+数据许可使用权（对应数据副本）”的数据基础产权制度；建立“源头确权登记+全量流通图谱”的数据流通全程在线化受控机制，支撑数据要素的规模化可信流通；引导市场建立“单次低价×多次复用”的数据收益长效机制。

3.3.2 技术系统搭建

实数融合是核心理念和基础框架，系统论与大模型是方法论支撑，“在全局性空间可信动态收敛”是破局思路，“全量流通图谱”与“数据流通大模型”是关键技术创新，全国“一个数据空间”和“一个基础设施平台（多节点）”是最终实现形态。在具体的技术实现上，应从系统整体角度出发，对多种可信安全技术进行组合裁剪。

3.3.3 市场生态运营

设立一个全国性的可信数据空间平台运营方，并明确其为数据要素市场建设运行的第一责任主体。同时，为实现市场生态的可持续发展，空间运营方只提供数据流通基本服务，如数据接入、全程存证、计费结算、争议处理等共性支撑功能；而针对具体流通事件的增值服务，则由专门的数据服务商来提供。

综上所述，“基于实数融合的可信数据空间”是内嵌于数据要素市场的基础操作系统。它通过全国化数据流通基础设施体系保障数据能够可信、安全、高效地流通，从而支撑基于数据要素市场的数据动态质量评价与双循环价值模型从理论构想走向工程实践。

4 工程实现:“高质量数据集专区”

要想实现前述“市场定义质量、双循环驱动”的范式，必须依托可信数据空间的底层能力，设计清晰、可工程化的技术路径与运营机制。本节将系统阐述高质量数据集从资源发现、产品化封装、交易达成、产品交付再到激励反馈的全流程，重点说明如何利用可信数据空间的既有能力，并创新性地通过设立“高质量数据集专区”（以下简称“专区”）来承载。

4.1 基础能力依托与“专区”设立

可信数据空间作为数据要素流通的基础设施，其核心能力为高质量数据集的生态化建设提供了不可或缺的支撑。高质量数据集建设并非另起炉灶，而是在此架构之上进行功的能延伸与运营细化。

4.1.1 核心能力基础:全量流通图谱与数据流通大模型

可信数据空间的全量流通图谱作为全局性、不可篡改的数据关系记录系统，是动态质量评价的“事实基座”。它自动捕获数据资源从确权登记、入市交易、访问调用到应用反馈的全生命周期事件。图谱中汇聚的流通频率、关联组合、应用场景、效果反馈等真实可信数据，为量化评估数据资源的“市场活性”与“场景效用”提供了客观、连续的输入。

数据流通大模型[11]作为可信数据空间的智能核心中枢，承担两大关键任务：一是“可信管控”，即通过实时监控与智能分析全量流通图谱，自动识别流通风险、执行合规策略，确保数据在复杂多边交互中的可信、安全与可审计；二是“价值发现”，即深度挖掘数据间的潜在关联与组合规律，主动识别高价值数据资源、预测其应用效能，并智能推荐最优流通路径与融合方案，从而驱动数据要素实现精准配置与价值最大化。

4.1.2 运营载体创新:设立“高质量数据集专区”

为清晰区隔双循环并实现专业化运营，需要在统一的可信数据空间基础设施之上，设立独立的“高质量数据集专区”。

（1）功能定位：专区是副循环（AI训练服务循环）的专门运营平台。它不替代主循环广泛的要素流通市场，而是作为其价值溢出的承接与转化枢纽。

（2）数据来源：专区内产品化的原始材料，并非直接来自原始数据资源提供方，而是主循环中已被市场验证的高质量数据资源。这些资源经由全量流通图谱标记和数据流通大模型推荐，具备明确的“高流通活性”或“高效用历史”标签。

（3）关系界定：主循环是价值创造与发现的广阔市场，解决实体经济赋能问题；专区是价值提炼与标准化输出的专业工场，服务AI模型训练。两者通过可信数据空间的底层能力贯通，形成“市场筛选、专区精加工”的协同流水线。

基于可信数据空间架构的“高质量数据集专区”示意图如**图2**所示。

4.2 高质量数据集建设:主体与流程

将经过市场验证的高价值数据资源转化为可供AI模型训练使用的高质量数据集产品，是一个涉及多角色、多技术的专业化过程。

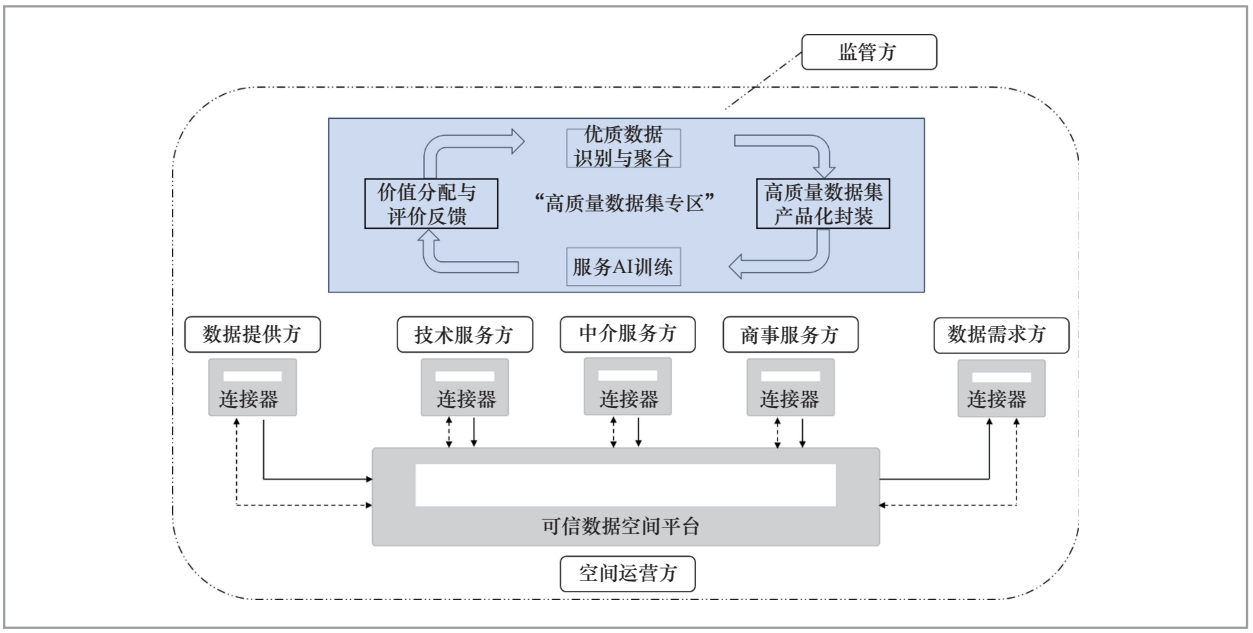

图2 基于可信数据空间架构的“高质量数据集专区”示意图

4.2.1 产品化主体：数据集产品专业服务商

高质量数据集的生产者不是原始数据资源提供方，而是引入的第三方专业数据服务商。他们扮演“数据精炼师”和“产品经理”的角色，具有如下职能。

（1）需求对接：与AI研发机构深入沟通，明确模型训练的具体目标、数据格式、质量要求与合规边界。

（2）资源遴选与聚合：基于数据流通大模型的推荐，从“专区”准入的资源池中，挑选出一批相关性高、质量历史优的数据资源。

（3）深度加工与封装：对遴选出的数据资源进行深度加工，包括任务对齐的精细标注、多源数据的语义对齐与融合、复杂场景下的数据增强、生成符合伦理与安全的合成数据，以及严格的隐私保护再处理（如差分隐私、k-匿名等）。

（4）产品标准化：将加工后的数据封装成标准化的数据集产品，提供完整的数据说明书，包含数据规模、分布、标注规范、潜在偏差、使用许可及更新计划等。

4.2.2 交付、结算与反馈闭环

（1）交易交付：加工完成的数据集产品在专区上架。AI研发机构以订阅制或项目制方式获取数据集使用权。具体交付方式包括：①对于通用大模型预训练需求，通过“有限白名单”等严格审核程序后允许离线拷贝；②对于模型微调、检索增强生成等更为普遍的需求，通过安全数据沙箱、隐私增强计算等形式进行在线流通（与可信数据空间主循环的方式一致）。

（2）计费结算：鼓励采用“基础许可费+效果分成”的复合模式。基础许可费覆盖数据集获取与加工处理成本；效果分成条款可约定，若使用该数据集训练的模型在特定商业场景中取得收入，则数据集服务商及数据资源提供方可按约定比例获得分成。结算由可信数据空间底层的数字

合约自动执行，计费依赖流通图谱中记录的应用验证等信息。

（3）训练效果反馈：AI研发机构在使用数据集完成模型训练和评估后，有义务（可通过合约激励）向专区提交对该数据集产品的综合评价报告（如在不同测试集上的表现等）。这些反馈信息将作为对该数据集产品持续评价的基础性依据，面向专区全局公开可见；同时反向更新数据流通大模型的知识，优化未来的智能匹配与资源推荐。

4.3 一个典型示例

为直观展示上述路径，以下结合一个简化的工业设备故障预测场景，描述高质量数据集“Spark-Industrial-Fault-Data V1.0”的构建流程。

步骤1：价值发现与资源准入（在主循环）

多家工业企业将脱敏后的设备传感器时序数据、维保记录作为数据资源，在可信数据空间主循环市场流通，用于设备健康管理服务。

经过一段时间的流通应用，全国流通图谱显示，来自A、B、C三家企业的数据在“泵机异常预测”场景中，被频繁组合调用，且用户反馈的预测准确率提升显著（均>15%）。数据流通大模型将这些资源标记为“高效用组合”，并推荐其进入“高质量数据集专区”资源池。

步骤2：产品化封装与供需匹配（在“专区”）

某专注工业AI的数据服务商在“专区”看到该推荐，并接收到某AI公司“智研模型”开发通用工业故障预测大模型的需求。

该数据服务商基于用户需求，利用隐私计算技术对A、B、C的资源进行联合分析，设计“多机型泵机故障图谱”构建方案。在保障各原始数据资源许可约束下，进行任务对齐的标注（如统一故障分类体系）、数据融合与增强，最终生成一个包含百万条样本的标准化数据集产品。

步骤3：交付、应用与反馈（闭环）

“智研模型”以“项目制+未来效果分成”方式订阅该数据集，在安全环境中用于模型微调。

训练后的模型在为客户提供故障预测服务中获得商业收入。根据智能合约，收入的一部分自动分成给数据服务商和原始数据资源提供方A、B、C。

“智研模型”向“专区”提交该数据集在内部基准测试上的提升报告。“专区”将此反馈记录于该产品的评价中，并用于优化数据流通大模型对未来工业数据资源的价值判断。

由以上流程可知，高质量数据集的建设不再是孤立的、高成本的预研型项目，而是植根于广阔数据要素市场、由市场验证驱动、由专业服务商执行、具备可持续商业闭环的生态化过程。

5 可持续商业模式与协同激励

第4节所阐述的技术路径，为高质量数据集的生态化建设提供了工程实现范例。然而，该范例能否成功且持久运行，关键在于能否构建一个与“双循环”价值模型精准匹配、能激励多方主体持续参与并合理分享价值的商业模式与治理体系。为此，本节旨在设计这一可持续的经济与社会契约。

5.1 价值创造与分配框架

“高质量数据集专区”商业模式的核心

在于承认并量化数据要素在从原始资源到AI模型价值增值全链条中各参与方的贡献，并据此进行公平、透明、自动化的价值分配。其参与方主要包括：

（1）数据资源提供方：位于主循环，是数据价值链的起点。其贡献在于提供经市场验证有效的高质量数据资源，并许可将数据资源纳入副循环进行产品化开发。

（2）数据集专业服务商：位于副循环，是数据集价值提炼与产品化的核心。其贡献在于将分散的、原始的数据资源，通过专业的知识、技术与劳动力，转化为可直接用于AI训练的高标准化、高附加值数据集产品。

（3）AI模型研发机构：位于副循环，是数据集产品最终价值实现的关键。其贡献在于利用数据集训练出具有强大能力的模型，通过模型服务创造商业价值或社会效益，并向上游支付高质量数据集的使用费用。

（4）专区运营方：位于副循环，作为专门的高质量数据集生态维护者，其贡献在于提供标准规则、推广运营及争议调处等应用层面的公共服务。其角色有别于可信数据空间运营方所扮演的底层基础设施提供者角色。

（5）AI模型应用方（实体经济中的广大主体）：位于主循环，既是AI模型的应用者，也是数据（集）赋能实体经济的最终价值创造环节。其贡献在于作为使用者付费，并提供对于数据和AI模型的使用反馈。

基于此，进一步提出“源头分成+服务增值+生态激励”的复合收益分配模型。该模型将数据集产品的总收益划分为三个部分，暂按“5∶3∶2”比例进行收益分配，具体如下。

（1）源头数据分成池（占比50%）：此部分收益直接回馈给数据资源提供方，内部分成比例根据其数据资源在最终数据集产品中的“贡献度权重”确定。该权重并非简单依据数据量大小而定，而是由数据流通大模型综合评估得出，主要考量因素包括：该资源在构成最终数据集任务中的不可替代性、数据稀缺性，以及更新维护活跃度等。

（2）数据服务增值池（占比30%）：此部分收益归属数据集专业服务商，是对其知识、技术与劳务投入的直接回报。收益与服务商的产品质量、市场口碑、售后服务（如数据更新、技术咨询等）紧密挂钩，激励其不断提升产品化能力与服务水平。

（3）运营与生态激励池（占比20%）：此部分由专区运营方支配。其中一部分（如5%）用于覆盖专区运营方的运行成本；另一部分（如15%）设立为“生态激励基金”，用于奖励对生态有特殊贡献的行为，例如，提供珍贵但小规模数据的“长尾”提供方、提交优质数据标注修正报告的AI研发机构，或成功推荐高价值数据组合方案的第三方开发者等。

5.2 基于流通贡献的激励算法

为将上述分配框架自动化、公正地执行，需要设计一套基于智能合约和链上可验证数据的激励算法。该算法的核心输入是可信数据空间全量流通图谱中记录的多维贡献证据，输出是面向各参与方的动态收益分配。算法的核心是量化贡献、自动结算。

5.2.1 对数据资源提供方的动态分成

设某数据集产品由n个原始数据资源加工而成。在结算周期内，获得的总收益在向某一数据资源分配时，需综合考虑该

数据资源在主循环的独立流通效用分数、与产品中其他资源的协同增益系数，以及其数据量占比、质量评分等。

5.2.2 对数据集专业服务商的绩效奖励

数据集专业服务商对某数据集产品的收益分配需考虑该产品在相关绩效指标上的表现，如用户平均评分、续订率、根据用户反馈完成重大改进的次数等。在保障专业服务商基础收益的同时，通过绩效奖励引导其关注产品长期质量和用户满意度。

5.2.3 对AI模型研发机构的反馈激励与效果分成

为鼓励AI模型研发机构提供高质量的模型训练效果反馈，并分享模型成功带来的增值，可设计以下机制。

（1）反馈奖励：研发机构提交经专区运营方验证有效的基准测试报告或改进建议后，可从“生态激励基金”中获得一次性现金奖励或积分奖励。

（2）效果分成合约：对于明确用于特定商业化场景（如医疗影像辅助诊断）的数据集采购，鼓励双方签订“基础许可费+后端分成”的数字合约。模型研发机构使用数据集产品时只需支付基础许可费用；当训练出的模型在具体场景中产生商业化收益时，模型研发机构按约定比例向数据集服务商与数据资源提供方分成。

5.3 “社区制”运营与生态治理

前述直接货币化分配模型为高质量数据集建设提供了基础的经济激励。然而，必须清晰认识到，AI模型特别是基础通用大模型的价值具有极强的正外部性（类似于基础科研），其对全社会生产效率的提升和创新的促进，远非模型研发主体落地变现的直接收益能够完全覆盖。若完全依赖其有限的商业收入反哺整条上游数据集价值链，生态在长期规模扩张与业务活力提升层面将陷入发展瓶颈。因此，必须为“高质量数据集专区”注入更强的公共产品属性与社区共创精神。可借鉴成熟开源社区的治理智慧，构建可持续的“社区制”运营模式。

首先，应锚定“专区”的公益属性，将核心运营权交由具备公信力的公益性机构承担。该机构不以利润最大化为经营目标，核心职能是维护生态规则、保障技术中立、促进广泛参与和确保长期存续。其主要收入来源应摒弃高额交易抽成模式，主要依靠覆盖基本运营成本的适度服务费、政府专项支持及社会捐赠。

其次，社区的重大规则变更、重大项目立项、重大争议裁决等，不得由运营方单方面决定，而应通过基于贡献积分的治理代币模型，由社区成员共同投票决策。该设计确保了生态的发展方向符合大多数贡献者的诉求，强化参与主体的归属感与主人翁意识，使治理结构更加民主、透明和富有韧性。

再次，为在直接货币激励之外激发更广泛、更深入的非货币化贡献，应设计一套精细化的“社区贡献积分”系统。该积分代表参与者在生态内的信誉与贡献度，通过算法自动发放，并支持内部流通与兑换。①积分获取：参与者可通过多种行为赚取积分，包括但不限于提供并维护高质量数据资源、开发或优化数据处理工具代码等。②积分流通与消耗：积分形成专区内部的“软通货”体系。其核心消耗场景包括：高积分参与者在获取紧俏数据集使用权限、参与前沿合作项目时可享有优先权；可兑换由社区运营方或合作方提供的

增值服务，如额外的云算力配额等。③关键创新：连接AI模型服务，社区参与者积累的贡献积分，可按一定规则兑换为合作大模型的优先体验权、更高的应用程序编程接口（application programming interface，API）调用配额、定制化模型微调服务等。

6 讨论与展望

6.1 理论贡献

本文提出的“基于数据要素市场的动态质量定义与双循环价值模型”，其核心理论贡献在于为高质量数据集建设提供了一套系统论指导下的生态化演进范式。它超越了当前主流的“工程化”与“项目制”思维，实现了三个关键的理论跃迁。

首先，实现了数据质量观的认知跃迁。传统观点将数据质量视为数据对象内在的、可独立测量的静态属性，本文则将其重新定义为一种“关系性”和“过程性”的涌现属性。质量并非预先嵌入数据之中，而是在数据与具体业务场景、多元市场主体的动态交互过程中持续生成和演化的。这一观点将质量评价从对“物”的测量，转向对“物在关系网络中所起作用”的评估，更符合数据作为高流动新型要素的本质。

其次，实现了数据价值论的路径跃迁。通过构建“实体经济赋能为主、AI训练服务为辅”的双循环模型，本文将数据价值实现从单一的、线性的“研发驱动”路径，重构为复合的、网络化的“市场驱动”路径。这从根本上扭转了数据价值依附于AI模型研发预算的被动局面，将其锚定在更为广阔和坚实的实体经济数字化需求上，为数据要素的价值创造与价值捕获提供了符合经济学原理的可持续逻辑。

最后，实现了技术系统观的范式跃迁。本文没有将技术（如时空码、数联网、流通大模型等）视为孤立工具的堆砌，而是将其置于“可信数据空间”的整体框架下，阐释了它们如何通过协同作用来支撑一个动态、可信、智能的数据要素市场。这体现了一种“系统设计”思维：技术组件是服务于“市场定义质量、生态化演进”这一核心范式要求的使能器，其价值和意义在系统整体功能中得到完整诠释。

6.2 实践意义

本文的研究成果为不同实践主体提供了清晰、可操作的行动指南。

（1）对数据资源提供方而言，本文指明了从“数据囤积者”或“一次性数据卖家”向“数据要素持续运营方”转型的方向。数据资源提供方的核心任务不再是完成孤立的数据集项目，而是将自身数据资源以标准化、可信的形式接入要素市场，并通过设计合理的收益模型，在数据的长期流通与应用中持续获取价值分成。

（2）对数据集专业服务商而言，本文给出了其角色从一次性、项目制“外包方”向可持续的“数据产品开发商”与“生态价值贡献者”的升级。数据集专业服务商的核心竞争力在于构建可复用的数据集加工产线、深入理解垂直领域知识，并通过“服务增值+绩效分成”等模式，在数据要素的长期价值循环中分享收益。

（3）对数据空间运营方和高质量数据集专区运营方而言，本文明确了两者的定位、分工与协同关系：数据空间运营方是面向全社会数据流通的基础设施运营主体，是具有一定公益属性、接受政府强监管的商业实体；高质量数据集专区运营方则定位为纯公益机构，面向高质量数据集这一

专门领域，在可信数据空间基础设施支撑下进行“社区制”运营。

（4）对AI模型研发机构而言，本文提供了破解高质量数据获取困境的新思路。AI模型研发机构应积极拥抱数据要素市场，从被动采购“成品数据集”转向“开放协作、主动参与生态”，通过签订“基础许可+效果分成”等新型合约降低前期成本、绑定长期利益；同时应积极将自身视为数据价值循环的关键反馈节点与价值共创方。

6.3 研究展望

基于当前研究，未来的研究可沿以下几个方向深化探索。

（1）法律与经济学交叉研究：深入探讨数据产权分置背景下，基于流通贡献的价值分配理论，及与之相适应的会计核算、税收征管等制度创新。

（2）数据流通大模型的算法创新：专注于开发更精准的数据价值预测模型、更高效的组合推荐算法，以及能够理解复杂业务语义的匹配模型，提升市场运行的智能化水平。

（3）行业试点与案例库的构建：在智慧医疗、普惠金融、智能制造等条件成熟的领域开展先行先试，积累成功的商业模式和治理经验，形成可复制推广的标杆案例库。

7 结束语

本文围绕高质量数据集建设面临的商业不可持续、评价标准脱节、与实体经济赋能割裂等现实困境，提出并论证了一个系统性解决方案：将高质量数据集建设深度融入数据要素市场体系，构建以“实体经济赋能”为主渠道、“AI训练服务”为副渠道的数据要素双循环价值模型，推动数据质量由市场化流通与应用反馈来动态定义。这一路径的实现，依赖于“基于实数融合的可信数据空间”所提供的架构支撑，和与之匹配的精准激励与协同治理框架。

本文是对高质量数据集建设范式的初步探索，本文所提方案从理论构想到规模化实践仍面临诸多复杂挑战，谨以此文抛砖引玉，期待能够引发业界与学界更深入的讨论、批评与实践校验。

参考文献：

[1] 中国信息通信研究院，国家数据发展研究院，中国电子技术标准化研究院，等. 高质量数据集建设指引[EB].2025.
CAICT, National Academy of Data Development, China Electronics Standardization Institute, et al. Guidelines for the construction of high-quality dataset [EB]. 2025.

[2] 林镇阳，吴江，胡鑫，等. 数据要素市场中高质量数据集评价指标体系建设研究[J]. 信息资源管理学报，2025，15(6)：52-66.
Lin Z Y, Wu J, Hu X, et al. Construction of an evaluation indicator system for high-quality dataset in the data element market[J]. Journal of Information Resources Management, 2025, 15(6): 52-66.

[3] 王闯，智佳琦. 推动AI高质量数据集建设促进数据资源开发利用[J]. 数字经济，2024(12)：22-25.
Wang C, Zhi J Q. Promote the construction of AI high-quality data sets and promote the development and utilization

of data resources[J]. Digital Economy, 2024(12): 22-25.

[4] 程乐. 我国高质量场景数据集的供给现状与发展策略[J]. 人民论坛, 2025(5): 68-72.
Cheng L. Supply status and development strategies of high-quality scenario dataset in China[J]. People's Tribune, 2025(5): 68-72.

[5] 姜春宇, 白玉真, 刘渊, 等. 构建企业级人工智能高质量数据集: 方法与路径[J]. 大数据, 2025, 11(6): 47-56.
Jiang C Y, Bai Y Z, Liu Y, et al. Building high-quality dataset for enterprise-level artificial intelligence: methods and pathways[J]. Big Data Research, 2025, 11(6): 47-56.

[6] 林镇阳, 胡鑫, 郭明军, 等. 可信数据空间协同治理下的高质量数据集建设与长效运营路径[J]. 图书情报知识, 2025, 42(5): 19-30.
Lin Z Y, Hu X, Guo M J, et al. The construction and long-term operation path of high-quality dataset under collaborative governance in trusted data spaces[J]. Document, Informaiton & Knowledge, 2025, 42(5): 19-30.

[7] 胡晓女, 李涛, 李姗姗. 人工智能大语言模型数据集现状和充实对策研究[J]. 大数据, 2025, 11(6): 57-71.
Hu X N, Li T, Li S S. Research on the status of large language model data set of artificial intelligence and enriching countermeasure[J]. Big Data Research, 2025, 11(6): 57-71.

[8] 张文娟, 邓辉, 艾政阳, 等. 我国AI大模型数据集建设发展刍议[J]. 人工智能, 2024(3): 85-95.
Zhang W J, Deng H, Ai Z Y, et al. On the construction and development of AI large model dataset in China[J]. AI-View, 2024(3): 85-95.

[9] 李林, 任伏虎, 罗超然, 等. 面向数据要素化的"实数融合"框架: 概念、必要性与系统实现[J]. 通信企业管理, 2026(3): 15-21.
Li L, Ren F H, Luo C R, et al. "Real number fusion" framework for data factorization: concept, necessity and system realization[J]. C-Enterprise Management, 2026(3): 15-21.

[10] 李林, 任伏虎, 蔡华谦, 等. 基于实数融合的全国化数据流通基础设施体系思考[J]. 信息通信技术与政策, 2025, 51(4): 15-27.
Li L, Ren F H, Cai H Q, et al. Thoughts on the system of national data circulation infrastructure based on the fusion of reality and data[J]. Information and Communications Technology and Policy, 2025, 51(4): 15-27.

[11] 李林,任伏虎,王宇奇,等. 基于实数融合的可信数据空间构建[J].大数据,2026:1-26.
Li L, Ren F H, Wang Y Q, et al. Construction of trusted data space based on real-number integration[J]. Big Data Research, 2026: 1-26

[12] 李林, 任伏虎, 蔡华谦, 等. 基于时空码和数联网技术的新型"可信数据空间"体系构想[J]. 信息通信技术与政策, 2024, 50(6): 89-96.
Li L, Ren F H, Cai H Q, et al. Conception of new trusted data matrix system based on spatio-temporal coding and internet of data technology[J]. Information and Communications Technology and Policy, 2024, 50(6): 89-96.

[13] 李林, 程承旗, 任伏虎. 北斗网格码: 数字孪生城市CIM时空网格框架[J]. 信息通信技术与政策, 2021, 47(11): 1-5.
Li L, Cheng C Q, Ren F H. Beidou Grid Code: CIM spatio-temporal grid framework for digital twin cities[J]. Information and Communications Technology and Policy, 2021, 47(11): 1-5.

[14] 罗超然, 马郓, 景翔, 等. 数据空间基础设施的技术挑战及数联网解决方案[J]. 大数据,

2023, 9(2): 110-121.
Luo C R, Ma Y, Jing X, et al. Internet of data: a solution for dataspace infrastructure and its technical challenges[J]. Big Data Research, 2023, 9(2): 110-121.
[15] 黄罡, 罗超然, 马郓, 等. 数据基础设施关键技术发展现状与挑战[J]. 科技纵览, 2023(12): 68-71.
Huang G, Luo C R, Ma Y, et al. Development status and challenges of key technologies in data infrastructure[J]. IEEE Spectrum, 2023 (12): 68-71
[16] 颜松远. 机器学习理论及应用[J]. 计算机工程与科学, 2012, 34(9): F0003.
Yan S Y. Machine learning theory and its application[J]. Computer Engineering & Science, 2012, 34(9): F0003.
[17] Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models[PP/OL]. V3. arXiv (2022-07-12)[2026-04-07]. arXiv: 2108.07258.
[18] 国家数据局. 可信数据空间发展行动计划(2024—2028年)[EB]. 2024.
National Data Administration. Trusted data space development action plan (2024-2028) [EB].2024.

作者简介

张群洪（1977– ），男，博士，数字福州集团有限公司董事长、高级工程师、高级经济师，主要从事数字经济与数据要素化、大数据与人工智能等方面工作。

李林（1980– ），男，博士，闽都创新实验室高级工程师、高级经济师，福建省数智双碳创新研究院院长，上海临港北京大学国际科技创新中心研究员，主要从事时空大数据与数字经济、数据要素与可信数据空间等方面工作。

杨堃（1990– ），男，现为福州数据集团有限公司总经理，主要从事数字经济与数据要素流通方面工作。

陈爽（1986– ），男，现为福州市电子信息集团有限公司总经理助理，主要从事人工智能应用技术方面工作。

张彬斌（1979– ），男，硕士，福州数据集团有限公司工程师，主要从事数据要素流通、数联网与可信数据空间等方面工作。

叶自燊（1990– ），男，福建省数智双碳创新研究院副研究员，主要从事数据要素流通与数据安全技术等方面工作。

收稿日期: XXXX-XX-XX

通信作者: 李林，leen999@126.com

基金项目: 福建省技术创新重点攻关及产业化项目“面向数据要素流通的数据资产可信管理系统关键技术研发及产业化”（闽工信函软件〔2025〕587号）；2025年度福州市人工智能“揭榜挂帅”科技重大项目（No.2025-ZD-036）

Foundation Items: Fujian Provincial Software Industry Key Technology Innovation Breakthrough and Industrialization Project (Mingongxinhan Ruanjian〔2025〕No. 587), 2025 Fuzhou “Jiebang Guashuai” Major Science and Technology Project for Artificial Intelligence (No.2025-ZD-036)