高质量数据集产品的形态和生产流程研究

杨琳[1,2]，朱扬勇[3]
1. 上海市大数据中心，上海 200003；
2. 华东师范大学数据科学与工程学院，上海 200062；
3. 上海数据研究院有限公司，上海 200120

摘要
高质量数据集决定了人工智能模型的训练效果。高质量数据集缺乏统一标准形态和质量可控的流程化生产方法，导致其供给不足、流通不畅，已成为制约人工智能发展应用的因素之一。本文从数据产品的角度，提出高质量数据集产品的五元组形态，以全链路技术能力为支撑，设计高质量数据集产品的生产流程，提出面向产品需求的全生产流程质量管控方法，为高质量数据集产品的大规模生产、流通提供理论基础和可行方案。

关键词
高质量数据集；数据产品；形态；生产流程；数据质量

中图分类号：TP399 文献标志码：A doi：10.11959/j.issn.2096-0271.2026036

Research on the form and production process of high-quality dataset products

Yang Lin[1,2], Zhu Yangyong[3]
1. Shanghai Municipal Big Data Center, Shanghai 200003, China
2. School of Data Science and Engineering, East China Normal University, Shanghai 200062, China
3. Shanghai Data Research Institute Co. , Ltd. , Shanghai 200120, China

Abstract
High-quality datasets determine the training performance of artificial intelligence models. The lack of a unified standard form and a quality-controllable, process-based production method for high-quality datasets has led to their insufficient supply and inefficient circulation, which has become a bottleneck restricting the development and application of artificial intelligence. From the perspective of data products, this paper proposes a five-tuple form for high-quality dataset products. Supported by full-link technical capabilities, we design a production process for such products and propose a product-oriented quality control method that covers the entire production chain. This work provides a theoretical basis and feasible solutions for the large-scale production and circulation of high-quality dataset products.

Key words
high-quality dataset, data product, form, production process, data quality

0 引言

高质量数据集建设已成为人工智能发展的核心工作之一，从国家顶层设计到地方创新实践，各行各业都在积极探索[1]。当前，针对高质量数据集的研究多数聚集于企业内部的建设，强调其作为模型训练“燃料”的技术属性，即经过采集、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集合[2]，而对于支撑其在数据市场流通的产品属性，如产品化封装、标准化交付、长效运营等关注不足。高质量数据集往往作为孤立的技术成果沉淀于单一项目或企业内部[3]，并不以产品的形式存在，难以实现规模化流通与应用，导致大量高质量数据集重复建设，并且标准不统一，训练效果难以保障。

随着人工智能的快速发展，组织内部自产自用的高质量数据集越来越难以满足模型训练需求，面临“有数难用、有市无货”的困境[4]。将高质量数据集从分散的、项目制的技术成果转变为能在数据市场中稳定流通与交易的数据产品，是解决这一问题的途径。其本质是将高质量数据集建设从“项目化定制”转向“产品化生产”。

本文将高质量数据集作为一种数据产品，研究产品的形态、生产流程和质量管控方法，主要贡献如下：①系统分析了高质量数据集产品应具备的组成部分，提出高质量数据集产品IDQCI（identity，dataset，quality，compliance，instruction）的五元组形态；②设计了高质量数据集产品的生产流程，明确六大生产环节和主要任务；③提出了面向产品需求的全生产流程质量管控方法，实现高质量数据集产品质量的可控可评估。

1 相关研究

数据产品的界定经历了一个较长的研究过程[5-11]。例如Pei[7]从应用角度将数据产品定义为从数据集派生出的用于产品和信息服务的数据集；黄丽华等[12]从数据流通市场建设角度、熊赟等[13]从数据产品流通监管角度分别对数据产品进行了研究。朱扬勇[14]针对数据产品给出明确定义：数据产品是指满足规范条件、可以独立使用、交易流通的数据集，具体包括数据产品唯一、数据对象完整、版权确定和数据产品可用等。数据产品由可用的数据集和对应的数据产品说明两部分组成。本文后续讨论均采用此数据产品定义。

（1）数据产品特征研究

数据产品应具备质量可评估、价值可流通、合规可监管的特征。

• 在质量可评估方面，美国麻省理工学院Wang等[15]将数据质量定义为“数据适合数据消费者的使用”，并提出数据质量与不同人员在具体环境下的“使用的适合性”程度正相关。林镇阳等[16-17]扩展了传统数据质量评估的固有属性指标，面向高质量数据集构建了包含数据规模、内容、价值等多维度的综合评价体系。随着数据从自用到他用的需求变化，蔡莉等[18]提出数据产品的质量是满足用户需求和监管需求的程度，将数据产品作为可流通的商品，并构建了一个数据产品的质量体系。

• 在价值可流通方面，叶雅珍等[19]提出以“盒装数据”为标准化产品形态，解决了数据产品可计量、可辨识的基础问题，奠定了数据产品在市场上有效流通的基础。黄丽华等[12]从交易成本与电子市场的视角

切入，将数据产品划分为4类并匹配相应的交易模式，提出数据交易平台降低交易门槛、增强供需黏性的方法。

• 在合规可监管方面。熊赟等[13]以“盒装数据”为可辨识、可计量的标的，构建涵盖产品质量评价、权属管理、流通过程监管等全周期的数据产品流通监管体系，实现了从产品形态到流通管控的延伸。

（2）高质量数据集产品封装和价值化

叶雅珍等[19]将数据盒模型产品化，设计了以“时间+空间+内容”三维数据立方体为盒内数据、以登记证书等为盒外包装的标准化“盒装数据”形态，旨在解决数据要素市场的计量与计价难题。参照盒装数据这一数据产品形态，对待交付的高质量数据集产品进行封装，将实现产品的可定价、可流通、可监管。明确的生产、封装、定价、交易机制为数据集创造了价值实现通道，并能通过市场反馈驱动数据持续迭代。同时，高质量数据集产品的价值实现能更加有效地促进数商生态构建，明晰权责与收益分配，从而真正激活数据要素市场的供需循环。

（3）高质量数据集产品生产

任洪润等[20]将数据工厂定义为场外流式数据市场中的数据产品生产加工点，是实现数据持续、快速、可靠流通与监管的核心要件之一。参考标准化、流程化的数据工厂生产模式，能够实现高质量数据集产品的规模化、工业化生产，保障合规、稳定的数据产品供给。质量功能展开（quality function deployment，QFD）是面向市场的产品设计与开发的一种计划过程，是质量工程的核心技术[21]。在该理念的启发下，可构建从数据要素到数据产品的五元组变迁矩阵，分析高质量数据集产品形态与生产环节之间的关联关系，规划高质量数据集产品生产环节和质量管控节点。

2 产品形态描述和变迁矩阵设计

2.1 高质量数据集产品基本要求

高质量数据集产品与传统数据产品或服务相比，存在明显的独特性。在形态上，传统数据产品或服务常以数据接口、数据表或分析报告形式交付，高质量数据集产品除提供数据集或数据接口外，更强调附带应用场景、数据模态、元数据、服务等级协议等信息的产品说明书，以实现其作为产品的可理解、可信赖和可复用。在作用上，传统数据产品或服务侧重于提供数据本身或基于数据的查询分析结果，而高质量数据集的核心是提供用于训练和优化人工智能模型的原材料或半成品，其价值不在于直接提供答案，而在于赋能使用者（模型）自身获得或提升某种能力。因此，高质量数据集产品除应具备质量可评估、价值可流通、合规可监管特征之外，还需具备内容可理解的特征。

• 质量可评估性要求建立客观、量化的数据产品质量评价标准，具备可实施的评估方法。“产品质量”维度能够承载数据准确性、完整性、一致性等内在技术属性，以及产品质量需求符合度等方面的度量要求，直观展现质量水平。

• 价值可流通性要求数据产品在市场中能作为独立的、具有应用价值的交易标的物。“产品标识”维度为产品赋予唯一身份，确立其在市场流通中的主体地位，以“封装数据集”为产品价值的具体载体，具备明确的权属，共同实现高质量数据集产品可流通的能力。

• 合规可监管性要求数据产品符合国家法律法规与伦理规范要求，并接受监管部门的监督。“产品合规性”维度对数据来

源合法性、隐私处理、内容安全等方面进行审核和承诺，在保障合规性的同时，也是监管的直接依据。“产品说明书”中的过程记录与“产品标识”提供的溯源信息则为监管的可追溯性提供了支持。

• 内容可理解性要求数据产品具备清晰的内在结构解读指引。“产品说明书”提供元数据、使用示例与场景说明等数据内容信息，保障产品的核心内容能直接被人阅读并理解。

综上，针对高质量数据集产品的四大核心特征，可以从5个方面构建一个完整、闭环的产品化描述体系：产品标识确立可流通主体，封装数据集承载可流通实体，产品质量界定技术价值，产品合规性划定法律边界，产品说明书则贯通理解、评估与使用。

2.2 五元组产品形态

使用产品化描述体系对高质量数据集产品进行描述，能够将高质量数据集从离散、不精确的自然语言描述，转化为一种可理解、标准化、结构化的表达，从而为后续的流程化生产、自动化评估和市场化流通奠定语义基础。高质量数据集产品P的五元组形态可表达为$P=\{I_d,D,Q,C,I_{\text{ns}}\}$，具体说明如下。

（1）产品标识

产品标识是产品的“身份证”，包括产品名称、版本号、唯一ID号、生产者、创建/发布时间等，用于识别和追踪该数据集产品。唯一ID号需要有统一的编码规则，编码应有足够的位数长度，以保证有充足的编码资源。ID号采用可读性强的分段组合码编码，各分段之间以中位短横线连接，编码长度为26位字符，编码结构为“区域码–厂商码–品类码–行业码–生产日期–序列号–校验码”。具体编码规则说明见**表1**。

（2）封装数据集

封装数据集是产品的“核心实体”，是按照生产流程生产，并满足合规、质量等所有约束条件，以特定格式（如特定目录结构的文件包、数据库表集合、API）组织起来的最终数据集合。

（3）产品质量

产品质量是高质量数据集产品价值的保障，也是“高质量”的依据，以质量说明书的形式呈现。质量说明书中对数据集在标准化质量指标评测水平Q_{stand}、基准模型水平提升的程度Q_{model}，以及与数据产品质量模型一致性的程度Q_{consi} 3个方面分别

表1 ID号编码规则说明

分段	长度	说明
区域码	2位	参照国家标准《中华人民共和国行政区划代码》中的规定，取前两位明确省级区域
厂商码	5位	省级区域内获得高质量数据集产品生产授权的机构或数据工厂统一编码，可由字母和数字构成
品类码	1位	1代表通识类高质量数据集，2代表行业通识类高质量数据集，3代表行业专用类高质量数据集
行业码	5位	参照国家标准《国民经济行业分类》，明确产品的行业领域特征。当不涉及中类或小类时，中类或小类顺序码以*填充（当品类码为1时，本码段取值为00000）
生产日期	6位	按照YYYYMMDD数字编码
序列号	6位	每厂商每品类每日从000001开始递增计数
校验码	1位	

进行计算和说明，得到相应客观、可比的产品质量水平描述值。

（4）产品合规性

产品合规性为高质量数据集产品生产与流通过程中涉及的合法性、安全性及伦理要求提供证明，以合规说明书的形式呈现，主要包括数据来源合规性、隐私保护符合性、内容安全性、分发模式许可等，是产品获得市场准入资格和构建可信交易基础的前提。

（5）产品说明书

产品说明书为高质量数据集产品可理解、可复用提供保障，主要包括元数据结构，如数据的内容、时间、空间、格式、规模；封装形态，如标准API、离线数据包或SaaS服务；应用场景建议，如交通规划、防汛防台、精准医疗等；服务等级协议，如可用性、响应时间、更新频率等。

2.3 从数据要素到数据产品五元组的变迁

高质量数据集产品生产的过程，是伴随着五元组形态逐步明确的过程。高质量数据集产品五元组形态并非在单一生产环节完成，是在生产过程中逐步填充、迭代的，并在关键节点实现确定和固化。对五元组形态与生产环节的相关性程度（包括强正相关、正相关、弱相关、不相关4个等级）进行分析：每一个形态维度都至少与一个生产环节强正相关。产品封装确定了最终的五元组形态，验证了生产环节设计的必要性和完整性。具体分析见**表2**。

（1）产品标识在产品规划环节初步明确，在设计数据架构时，就会界定数据的来源、范围、关键属性，并规划产品唯一ID号和元数据框架等，为其赋予最初的身份信息，最终在产品封装环节确认产品标识相关信息。

（2）封装数据集按照产品规划环节设计的各项特征指标，在数据要素获取、清洗与标准化、数据标注与增强等所有生产环节中逐步生产成形，通过产品封装形成与产品说明书匹配的数据集实体。

（3）产品质量贯穿整个生产全流程。产品质量模型构建始于产品规划阶段，定义目标质量指标；质量控制分布于清洗与标准化、数据标注与增强、质量检测等各个生产环节，及时发现质量问题；产品封装环节确定最终的质量水平，形成质量说明书。

（4）产品合规性在数据要素获取环节确认，对采集到的数据对象进行合规性审核，包括是否得到购买、共享、交换等数据获取方式的授权，采集数据是否含有个人隐私、对国家安全造成重大危害的内容

表2 数据要素到数据产品五元组变迁框架

生产环节	产品标识	封装数据集	产品质量	产品合规性	产品说明书
产品规划	强正相关	正相关	强正相关	正相关	强正相关
数据要素获取	弱相关	正相关	正相关	强正相关	弱相关
清洗与标准化	弱相关	强正相关	强正相关	不相关	弱相关
数据标注与增强	弱相关	强正相关	强正相关	不相关	弱相关
质量检测	弱相关	强正相关	强正相关	不相关	弱相关
产品封装	正相关	强正相关	正相关	正相关	强正相关

等。采集到的数据通过合规性审核后，在产品封装环节生成合规声明。

（5）产品说明书在产品规划环节形成内容框架，完成数据要素获取、清洗与标准化、数据标注与增强、质量检测等各个生产环节后，在产品封装环节汇总所有信息成稿。

3 产品生产流程及质量管控方法设计

高质量数据集产品的生产是一个以合规为前提、以质量为基线、以五元组形态实现为目标的系统性过程，遵循“合规的数据才能用于生产、合格的数据才能出厂流通”原则。本文设计了产品规划、数据要素获取、清洗与标准化、数据标注与增强、质量检测、产品封装6个核心生产环节，并设置了相应的质量检测或合规审核节点，实现嵌套和多循环的“检测–反馈–优化”机制，形成从需求到交付的完整生产流程。各环节质量检测结果发挥双重作用：其一，作为判断本环节产出是否达到预设标准、能否流入下一环节的决策依据；其二，作为反映本环节生产过程质量水平的反馈信号，用于驱动该环节作业流程与方法的迭代优化。高质量数据集产品生产流程如**图**1所示。

3.1 产品规划

产品规划环节是高质量数据集产品生产的起点，其目标是将模糊的业务需求转化为明确的产品定义与质量基准，为后续生产活动提供执行依据。本环节的主要任务如下。第一，进行需求分析与定义，明确产品拟支撑的业务场景、目标用户，确定数据的模态构成、规模量级以及覆盖的行业领域等，形成清晰且无歧义的功能性与非功能性需求描述。第二，基于已明确的需求，构建量化的产品质量模型。该模型可依据国家或行业标准，将需求分解为

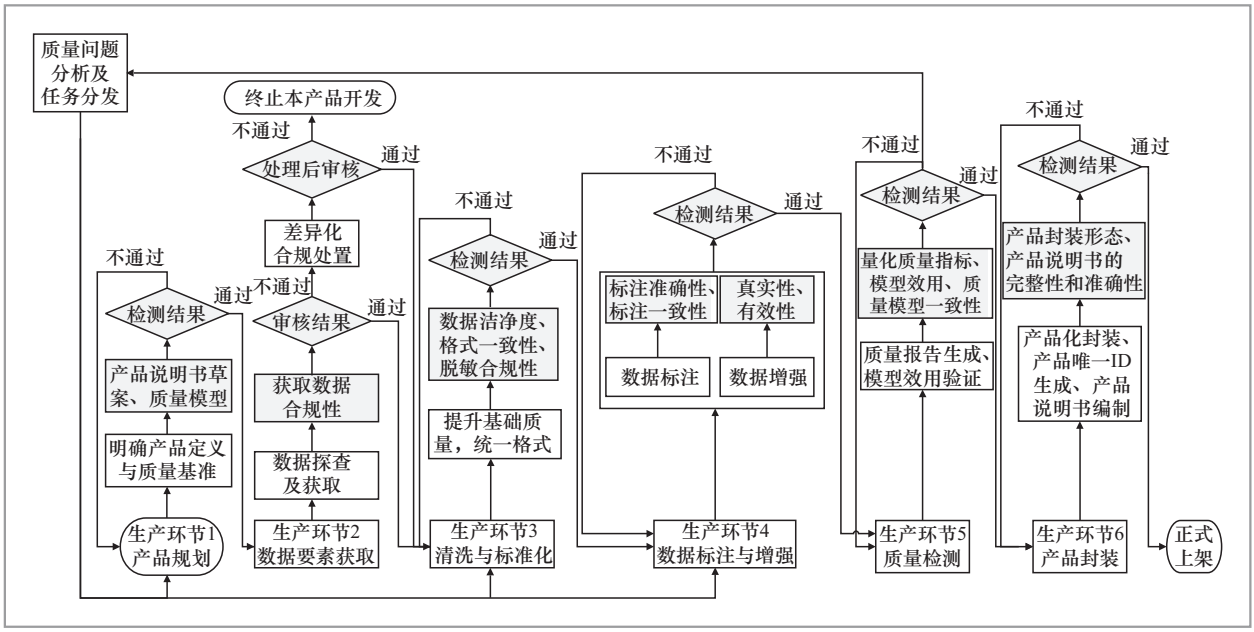

图1 高质量数据集产品生产流程

基础数据质量、内容语义质量、场景应用质量及安全合规要求等多个维度，并为每个维度设定可测量的具体指标与阈值。第三，完成产品初步设计，输出产品说明书草案。该说明书需涵盖产品元数据信息，包括数据来源、采集时间范围、数据规模与模态、标注规则与标签体系、预期的适用场景与局限性、授权范围等。

本环节主要对产品说明书草案和产品质量模型进行质量检测，检测重点包括需求定义的完整性、清晰性与可度量性，确保没有歧义且所有需求均可被后续设计响应；产品质量模型的完备性与科学性，检查指标维度是否覆盖所有关键质量属性，指标阈值设定是否合理且有依据。若检测发现问题，则需调整完善产品规划，直至其成为合格的生产输入依据。本环节主要涉及用例分析、质量功能部署等结构化需求工程技术，系统化地捕获、分析和满足优先级排序需求；采用质量指标体系建模技术，构建层次化、可量化、可操作的质量评价模型。

3.2 数据要素获取

数据要素获取环节的主要任务是执行多源异构数据采集，通过产品规划确定的公共API、传感设备、公开数据集、商业数据采购、内部系统导出等多个渠道，在确保授权合法有效的前提下进行数据收集与汇聚。

本环节主要对数据获取的合规性进行审核，审核重点包括数据来源合法性和授权合规性，确认数据提供方是否具备相应资质、数据采集行为与授权协议是否符合法律法规要求；数据内容安全性，筛查数据中是否含有涉及国家秘密、危害国家安全和社会稳定等法律明文禁止或违反公序良俗的内容；隐私泄露风险性，识别数据中是否包含未经处理的个人敏感信息，依据“最小必要”原则审视采集的数据是否超出了实现处理目的的最小范围，并评估其潜在风险等级。对于确认存在合规性风险的数据，要针对不同类型的不合规情形采取差异化的处置策略，在保障合规的前提下尽可能降低对数据集规模和质量的影响。若处置完成后再次审核仍然存在合规性风险，则本产品开发终止。

本环节主要涉及数据采集和合规性审核技术。多源适配采集技术能够实现对不同协议和接口的适配；实时数据流监控与边缘处理技术在数据源头实现实时采集、即时质量监控、实时合规筛查、边缘侧预处理、动态反馈与即时响应等功能；自动化合规扫描技术利用规则引擎与敏感信息识别模型对流入数据进行初步过滤与风险标记；构建内容安全指纹库，可实现对违法和不良信息的快速比对与拦截。

3.3 清洗与标准化

清洗与标准化环节对获取的原始数据进行首次加工，旨在提升其基础质量，消除原始数据中的杂质，并将其转化为格式统一、规整的中间产品，为后续的知识注入和价值提升奠定基础。本环节的主要任务包括：第一，进行数据清洗，运用规则与算法去除数据中的乱码、无关字符等噪声，处理异常值与缺失值，检测并删除重复记录；第二，实施数据标准化，统一数据的格式、编码、计量单位、时间戳格式等，确保数据在形式层面具备一致性；第三，对在上一环节识别出的敏感信息，依据规划中的隐私保护要求，进行脱敏、匿名化或伪名化处理；第四，系统性地记录并维护数据血缘信息，即追踪数据从源端

开始，历经每一步处理操作的起源、变换与移动历程，确保全过程可追溯。

本环节主要对经过清洗、标准化和脱敏处理后的数据进行质量检测。检测重点包括评估数据洁净度，即评估去重、去噪、异常值修正等操作的实际效果；验证格式与标准的一致性是否完全符合预设规范；审查脱敏处理的合规性与有效性，确保敏感信息得到恰当保护且未损害数据的可用性；确认数据血缘记录的完整性，保证所有处理步骤均有迹可循。检测结果直接决定数据能否进入下一环节。若未通过，需重新进行清洗与标准化，相关结果反馈用于优化清洗与标准化算法。本环节主要涉及基于规则、模式识别或统计方法的自动化数据清洗技术、数据标准化引擎、差分隐私与泛化等隐私保护脱敏技术，以及能够实现数据血缘关系分析和记录的数据血缘追踪技术。

3.4 数据标注与增强

数据标注与增强环节通过注入领域知识和扩充数据样本提升数据的信息密度、语义丰富度与多样性，从而直接增强其对模型训练的贡献。

（1）数据标注

数据标注的主要任务是根据产品规划中定义的详细标签体系，为清洗后的数据样本赋予准确的语义标签、分类信息或结构注释，包括为图像中的物体绘制边界框并分类、对语音进行文本转写、为文本标注实体与关系等。对于多模态数据，还需进行跨模态对齐，确保不同模态的数据在语义和时序上正确关联。本环节主要对标注过程产出的标签数据及对齐结果进行质量检测。检测通常采用抽样方式进行，重点评估标注准确性（与专家或“黄金标准”的一致性）和标注一致性（不同标注员或不同批次对同类样本标注结果的一致程度）；对于多模态对齐结果，需评估其关联的精确度。

（2）数据增强

数据增强的主要任务是通过算法手段，在保持数据核心语义不变的前提下对现有数据集进行变换或合成，以生成新的衍生样本，从而解决数据稀缺、类别不平衡或提升模型泛化能力的问题。具体包括基于变换的增强，如图像的旋转裁剪、颜色的随机调整、文本的同义词替换和回译，以及基于生成式模型的合成，如使用生成对抗网络(generative adversarial networks，GAN)、变分自编码器（variational autoencoder，VAE）或扩散模型生成符合真实分布的新数据。本环节主要对增强或合成后生成的新数据进行质量检测。检测重点是评估其真实性（是否符合领域逻辑）和有效性（是否提升了数据多样性而未引入有害偏差），必须确保新数据仍然保留了原有的核心语义信息。

数据标注与增强环节检测中发现的问题样本需返回重加工，同时检测统计结果（如标注一致率、错误类型分布）用于优化标注指南、培训标注人员或调整增强算法参数。本环节在数据标注方面主要涉及人机协同智能标注技术，例如利用预训练模型进行自动预标注以提升效率，结合主动学习策略筛选不确定性高的样本进行人工重点标注；采用基于置信度过滤与协同校正的算法识别和修正标签错误；多模态对齐与融合技术，涉及跨模态检索模型、共享表征学习等方法。在数据增强方面，主要涉及传统的数据增强方法库和基于生成式人工智能模型的生成式数据合成技术。

3.5 质量检测

质量检测是对基本成形的数据集产品进行全面、标准化的检验与验证，确保交付的产品符合既定的质量承诺。本环节的主要任务包括：第一，执行标准化质量指标评测，利用自动化工具，依据产品规划环节定义的产品质量模型对数据集的各项质量指标进行批量计算与评估，生成量化质量报告；第二，进行模型效用验证，将数据集应用于一个或多个基准模型的训练与测试，以模型性能的提升为数据质量在应用层面的验证；第三，可引入第三方独立评测或执行内部交叉验证，以增强质量评估结果的客观性与公信力。

质量检测是生产流程中质量管控的核心环节，完整实现了量化质量指标达标情况、模型效用达标情况、产品质量模型一致性3类质量指标的检测。本环节的检测具有“闸口”性质，若未通过，必须进行根因分析，定位问题的具体环节（如标注错误、覆盖不足、采集偏差），并将问题回溯至相应上游环节进行整改，检测结果也用于评估和优化整个质量检测体系本身的效率与有效性。本环节需要采用自动化质量评测工具链，集成针对不同质量指标的专用检测算子进行检测，使用评估与误差分析技术开展基于基准模型的性能测试。

3.6 产品封装

产品封装环节为通过最终质检的高质量数据集赋予商品形态，使其具备质量可评估、价值可流通、合规可监管、内容可理解的特性。本环节的主要任务包括：第一，进行产品化封装，根据市场需求和交付协议，将数据集打包成特定的形态，如提供标准化的API、发布为特定格式的数据包等；第二，生成产品唯一标识与完整溯源档案，利用可信技术为产品分配唯一ID编码，并将全生命周期的关键处理日志、质检报告等信息进行不可篡改的存证；第三，编制最终版的产品说明书，全面、准确地描述产品内容、规格、使用方法、限制条件等；第四，附上最终的法律与商业文本，包括最终版数据许可协议、合规声明、质量说明书等。

本环节主要对最终封装完成的数据产品包及其附属文档进行质量检测。检测重点包括验证封装形态的功能性、兼容性与性能；产品说明书与元数据的完整性、准确性，确保其真实反映产品状况，且符合相关格式标准；溯源信息完整性与可验证性；法律文本的完备性与无歧义性等。这是产品出厂前的最终检测，任何缺陷都将阻止产品发布，检测结果同时用于优化封装模板、元数据生成工具和文档编写规范。本环节需要结合API网关设计、数据序列化/压缩算法等，实现数据产品封装与交付；结合数字身份与可信存证技术，为产品创建唯一的数字身份并固化溯源信息；结合结构化元数据自动生成技术，从生产流水线各环节自动提取信息，按照标准格式组装成元数据文件；基于智能合约技术实现数据流通过程中数据许可协议中的相关条款等。

4 结束语

高质量数据集的产品化需要系统化、工程化体系的构建，以解决数据供给、数据流通、数据应用和数据价值实现等核心问题。高质量数据集产品的五元组形态描述、生产流程设计实施，能够将数据要素转化为质量可评估、价值可流通、合规可

监管、内容可理解的数据产品，为高质量数据集产品的大规模生产、流通提供理论基础和可行方案，赋能人工智能产业纵深发展。后续笔者将重点研究高质量数据集产品生产线建设、高质量数据集质量控制管理体系建设等。

参考文献：

[1] CCSA TC601大数据技术标准推进委员会. 高质量数据集实践指南(1.0)[R]. 2025.
CCSA TC601 Big Data Technology Standards Promotion Committee. Practice guide for high-quality datasets (Version 1.0)[R]. 2025.

[2] 中国电子工业标准化技术协会. TC609-5-2025-01高质量数据集建设指南[R]. 北京：中国电子工业标准化技术协会，2025.
China Electronics Standardization Association. TC609-5-2025-01 high-quality dataset-construction guide[R]. Beijing: China Electronics Standardization Association, 2025.

[3] 中电数据产业集团有限公司. 中央企业高质量数据集建设研究报告[R]. 2025.
China Electronics Data Industry Group Co., Ltd. Research report on high-quality dataset construction for central state-owned enterprises[R]. 2025.

[4] 中国信息通信研究院. 高质量数据集建设指引[R]. 2025.
China Academy of Information and Communications Technology. Guidance for high-quality dataset construction[R]. 2025.

[5] Loukides M K. What is data science[M]. Beijing: O'Reilly Media, 2011.

[6] Hazen B T, Boone C A, Ezell J D, et al. Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications[J]. International Journal of Production Economics, 2014, 154: 72-80.

[7] Pei J. Data pricing: from economics to data science[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2020: 3553-3554.

[8] Huang G Y, He J, Chi C H, et al. A data as a product model for future consumption of big stream data in clouds[C]//Proceedings of the 2015 IEEE International Conference on Services Computing. Piscataway: IEEE Press, 2015: 256-263.

[9] Bengfort B, Kim J. Data analytics with Hadoop: an introduction for data scientists[M]. Beijing: O'Reilly Media, 2016.

[10] Cao L B. Data science: challenges and directions[J]. Communications of the ACM, 2017, 60(8): 59-68.

[11] 朱扬勇，熊赟. 数据的经济活动及其所需要的权利[J]. 大数据，2020，6(6)：140-150.
Zhu Y Y, Xiong Y. The required authorization to the data-centric economic activities[J]. Big Data Research, 2020, 6(6): 140-150.

[12] 黄丽华，窦一凡，郭梦珂，等. 数据流通市场中数据产品的特性及其交易模式[J]. 大数据，2022，8(3)：1-14.
Huang L H, Dou Y F, Guo M K, et al. Features and transaction modes of data products in data markets[J]. Big Data Research, 2022, 8(3): 1-14.

[13] 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.

[14] 朱扬勇. 相似点集挖掘实验数据集[M]. 上海：上海科学技术出版社，2021：11.
Zhu Y Y. Similar point set mining experimental dataset[M]. Shanghai:

Shanghai Scientific and Technological Publishers, 2021: 11.

[15] 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.

[16] 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 datasets in the data element market[J]. Journal of Information Resources Management, 2025, 15(6): 52-66.

[17] 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 datasets for enterprise-level artificial intelligence: methods and pathways[J]. Big Data Research, 2025, 11(6): 47-56.

[18] 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 L, Zhu Y Y. From data quality to data products quality[J]. Big Data Research, 2022, 8(3): 26-39.

[19] 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.

[20] 任洪润, 朱扬勇. 数据管道模型: 场外流式数据市场形态探索[J]. 大数据, 2023, 9(3): 15-28.
Ren H R, Zhu Y Y. Data pipeline model: exploration of the over-the-counter form of streaming data[J]. Big Data Research, 2023, 9(3): 15-28.

[21] 刘鸿恩, 张列平. 质量功能展开(QFD)理论与方法: 研究进展综述[J]. 系统工程, 2000, 18(2): 1-6.
Liu H E, Zhang L P. Quality function deployment theories and methods: review on research progress[J]. Systems Engineering, 2000, 18(2): 1-6.

作者简介

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

朱扬勇（1963– ），复旦大学教授，上海数据研究院有限公司学术副院长，《大数据》期刊编委会副主任。数据科学家，在数据领域做出开创性工作，提出数据自治、数据财政、数据资产等概念和体系。在数据领域发表论文200多篇，持有高价值专利15件，出版学术著作17本。上海市数据科学重点实验室创始主任，担任多个政府顾问。主导编制了多个地方政府数据发展规划。主要研究方向为数据科学与数字经济。

收稿日期: 2026-01-31

通信作者: 朱扬勇, yyzhu@fudan.edu.cn