

我国高质量数据集建设现状

文 ◆ 浙江大学光华法学院教授、博导
浙江大学国际战略与法律研究院常务副院长
数字法治研究院首席专家

程乐

在新一轮数字化革命的推动下，数据成为继土地、劳动力、资本、技术之后的第五大生产要素，并在不同社会领域配合其可复制性、可增强性、可训练性、可互操作性等独有特征，实现跨行业价值释放。

然而，未经筛选整合的海量数据依然存在数据噪声、非对称性、低完整性等内生问题，难以通过可信利用转化形成可持续的数据价值。面对迥然相异的产业技术需求，数据可以通过特定的格式和结构加以集合进而发挥实质价值，这一概念亦被称作“数据集”。

在新质生产力驱动的全新技术业态下，以人工智能为代表的尖端科技正在对特定垂直场域中的“场景数据集”展现更强的上下文依赖性，医疗、交通、教育、金融等场景化的专业性数据集需求高速提升，特定应用场景的精细化对数据价值体系提出更高要求。此背景下，场景数据集的概念激活与高质量发展正密切关联我国数字产业经济的创新性配置、领域性转型与技术性突破。

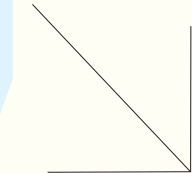
为实现场景数据价值效用的乘数倍增与充分释放，亟须打通我国关联领域目前在数据供给、数据流通、数据评估、数据标准、数据开放、数据共享等层面的堵点，以“场景化加工能力”与“多样化共享体系”两大要点共同构建具有国际化样板意义的高质量场景数据集。

一、高质量场景数据集的战略定位与价值释放

现代科技谱系中，数据集可能关联区块链、物联网、人工智能、自动化工程、高性能计算、地理信息系统等多端技术，作为中枢性资源要素发挥复合作用。而在医疗、工业等具体领域的技术应用中，数据集的价值转化则清晰展现“应用牵引”下的场景差异化需求。

例如，工业场景需要传感器数据等精度高、实时性强的数据集，医疗场景的数据集建构则更多关联隐私保护、数据保真度等要求。此背景下，面向具体场景的高质量数据集供给日趋关键，并对各国科技产业的战略性部署展现出重要价值。例如，欧盟委员会于2022年发布《高价值数据集实施法案》，明确环境数据、地理空间数据与地球观测数据等特定场景高价值数据集的供给标准，进一步完善面向工业、国防等重点场景的高质量数据集供给。

在多类应用情境中，人工智能对于高质量场景数据集的需求尤为突出。2024年《政府工作报告》首次将“人工智能+”上升至国家战略层面，明确提出深化大数据、人工智能等研发应用。近年来，以DeepSeek、ChatGPT、Gemini、通义千问、豆包为代表



的大语言模型在各社会领域展现了卓越的应用潜力与现实效果，此类生成式人工智能所采用的预训练方法便是通过海量数据的累积效应实现模型的智能涌现与能力迁移。

但一方面，预训练模式对于模型专业领域的性能提升效果欠佳，未经筛选的低质量数据容易导致专业性问题的“模型幻觉”（模型生成了不符合事实或毫无根据的信息）。另一方面，预训练所需的可用数据资源即将被用尽。在双重挑战的协同作用下，针对特定场景的高质量数据集成为人工智能模型训练的关键支撑，为现有数据局限下智能模型的训练路径优化提供转向可能。

其一，特定领域的高质量数据注入能够推动人工智能掌握对

应领域的关键特征与独有规律，通过垂直场景的上下文关系增强模型训练的精准性与针对性，实现高度专业化与个性化的人工智能解决方案与创新拓展。

其二，高质量场景数据集在模型微调与强化学习阶段的嵌入亦可抑制数据偏差与特殊场景下模型的表现失衡，基于目标行业的标准与规范提升模型的合规性、可信度与可解释性。

其三，OpenAI 推出的一系列推理模型正在数学、物理、化学等强逻辑专业领域展现专家级别的优异表现，此类后训练模型的调试与性能增强对于专业领域数据与反馈的需求更为迫切。2024 年末，OpenAI 宣布将开展强化微调（Reinforcement Fine-Tuning）的模型定制计划，通过特定领域小规模数据集的重点训练将通用模型转化为专业模型，并展示了运用此种方法推动 GPT o1 Mini 模型高水平推理罕见疾病成因的实例。

未来，更多专业维度的高质量场景数据集将助推人工智能于更广阔的重点领域展现“专家级”能力，实现模型性能与社会效率的场景化跃升。

二、我国高质量场景数据集的供给现状与发展困境

规范层面，我国早在 2020 年已认识到高质量数据集供给对于经济社会发展的战略价值，并发布《关于构建更加完善的要素市场化配置体制机制的意见》等系列政策。近年来，我国相继颁布《关于构建数据基础制度更好发挥数据要素作用的意见》（2022 年）《网络数据安全管理条例》（2024 年）等政策法规。

然而，相较欧盟在《公共部门信息复用指令》等开放数据法规基础上所明确的《关于开放数据和公共部门信息再利用指令》等规范及其落地效果，我国数据集的场景优势与场景化激活能力仍显不成熟，特别是在数据基础设施建设、公共数据开放制度、数据集共享生态等层面仍显不足。

实践层面，我国数据资源总量优势明显，多样化数据资源丰富，为高质量场景数据集的高速增长提供现实基础。据新华社消息，2024 年，全国数据市场交易规模预计超 1600 亿元，同比增长 30% 以上，其中场内市场数据交易（含备案交易）规模预计超 300 亿元，同比实现翻番。

根据全国数据资源调查工作组发布的《全国数据资源调查报告（2023年）》，2023年，全国数据生产总量达32.85ZB（泽字节），同比增长22.44%。

我国数据资源“产—存—算”的规模优势已基本形成。根据国际数据公司（IDC）的预测，中国“数据圈”（每年被创建、采集或是复制的数据集合）在2025年增至48.6ZB（泽字节），占全球27.8%，成为最大“数据圈”。同时，线上支付、电子商务、共享经济、电子政务、智慧医疗等多样化数字服务与数据应用场景在我国具有良好的社会接受度与广阔的应用前景。

得益于少数民族文化的包容特性与社会秩序的长期稳定，我国大多数民众对跨场景数据应用持开放态度，数字化需求趋于多样，这为数据赋能的技术突破与多领域数字创新的涌现提供了社会土壤。然而，我国高质量数据集供给的场景完善度不足，海量数据的潜能与多样化场景优势的潜能仍有待进一步释放。

其一，我国面向场景的有效数据供给有待强化，数据资源总量优势的价值转化尚待释放。在“原始数据—场景数据—可用场景数据—高质量场景数据集”的四阶段数据转化与价值递进中，作为中间产品的数据要素须经清洗、标注、分析等加工活动以实现价值创造。

此过程中，三方面因素使得我国在基于差异化场景需求实现数据价值的场景化增长方面存在一定难度。

一是数据加工过程中投入与回报不成比例所引发的动力匮乏问题。专业化数据的高质量标注需要庞大的资金投入且同时关联安全风险，这与数据应用的收益回报比例之间存在错位。

二是公共数据场景化加工的程度与紧迫性不足，难以发挥公共数据对于数据开发利用的引领作用与催化作用。从现有国家政策来看，我国工业、医疗、交通、气象等具体场景的数据供给能力尚未作为专门对象加以细化规定，领域差异化数据统归于公共数据范畴的做法难以匹配垂直领域的细化要求。

三是数据加工技术存在优化空间，专业人才供给难以充分满足需求，致使场景化加工效果欠佳。在数据标注自动化工程尚未成熟的阶段，将

杂乱无序的低价值数据在工业互联网等特定场景加工为标准化、目录化、高价值数据资源需要高级别的专业技术技能，人工智能训练与推理阶段涉及多模态数据、高维度数据、跨领域数据，对计算机科学技术需求更加严苛，这为我国数据加工的现代化场景转向带来一定挑战。

其二，受限于较低级别的数据开放度与流通度，我国高质量场景数据集的源头供给能力有待提升。首先，我国公共数据开放程度有限，关键主题高价值数据集的动态数据、应用编程接口（API）与数据开放许可证等标准尚不明晰，难以实现公共数据面向公众和企业的有序开放或强制开放。其次，我国多领域数据源很大程度上为大型数据平台所垄断，各行业数据普遍向平台呈现聚拢态势，根据全国数据资源调查工作组发布的《全国数据资源调查报告（2023年）》，大型平台企业平均数据交互量约为行业重点企业的200倍以上。但同时，大型平台企业常以保护个人信息、防范安全风险等理由拒绝共享，仅把数据用于自身场景构建。因此，大量中小型企业虽有

丰富的场景数据需求，却常处于“无数据可用”的困境。最后，在数据共享受限的背景下，我国数据交易亦存在供需失衡。《全国数据资源调查报告（2023）》显示，调研的27家交易所的数据产品中仅有17.9%实现交易。数据权属界定、数据交易定价等固有难题引发的数据交易制度滞后问题致使整体性数据产品成交率低迷，供需不匹配的现象广泛存在，数据交易机构供给水平难以满足场景数据需求。

其三，细分应用场景的技术标准仍有待完善，海量数据在垂直场景实现价值激活面临挑战。不同应用场景下，数据的格式、版本等特征性技术需求对应相异的数据标准与技术标准。例如，在柔性制造领域，数据需配合高频采样、实时监测的产业需求，在生产节拍精准同步等方面存在特定数据规范；在智能交通领域，位置、路况信息之外的视觉、雷达、激光测距等多样化数据融合及相应技术标准对高精度地图的更新则至关重要。

然而，我国当前主要着眼数据供给数量的增加，尚未针对场景深入开展数据集标准化技术工

程。此背景下，数据技术标准与架构的不统一增加了场景数据集成的复杂性，不同数据间的场景融合与预期场景的具体需求所关联的转换、加工与适配工作消耗过量资源，亦可能基于标准引发的数据误差降低场景数据供给的整体质量。

人工智能模型结构高速迭代的产业背景下，多模态模型所需图像、语音、文本、视频等数据形式的融合需求日趋关键，推理模型对高逻辑领域复杂问题的反馈数据需求也趋于迫切，多模态数据元模型拟定、元数据描述语言等新兴技术标准问题正快速更新，亟须有效回应。未来，细分场景下数据技术标准与架构的滞后效应可能进一步影响场景数据集在现代化技术需求下的流通与复用，阻碍数据流通、数据交易与数据价值增值的充分实现。

其四，现有规范体系下数据开放目录、数据开放主体、数据开放责任等配套制度在落地实施时面临一定阻碍，存在供给不积极、供需不匹配、共享不充分等问题。在实践中，真正实现开放的公共数据主要是各部门的政务数据，而此类数据难以直接转化为产业价值。而水电、通信、交通、医疗等重点场景数据则往往被拒绝开放，或限缩开放范围。

同时，在《中华人民共和国网络安全法》《中华人民共和国数据安全法》与《中华人民共和国个人信息保护法》这三大立法的实施进程中，数据控制合法性判断与数据流通利用的责任规则仍不明确，个人隐私与商业秘密难以在真正实现场景化数据开放的情境中得到充分保护。

此外，数据产权定义尚未统一，信息脱敏与数据加密等配套安全措施仍不完善，场景数据开放时法律责任与权益保护也因而趋向复杂化和模糊化。综合作用之下，即便我国近年已在相关法规中展现“推动按用途加大供给使用范围”等政策转向，但重点场景的数据确权与开放亦存在一定障碍，数据开放与数据安全的平衡性问题依然较难明确。

三、我国高质量场景数据集的多方共享与突破路径

其一，培育多方技术，强化高价值数据的现代化场景加工能力。现代化技术驱动的数字产业变革中，前沿科技对我国数据市场与场景化数据供给能力的重塑具有重要意义。首先，提升捕捉场景化需求的技术能

力，积极运用机器学习中的自然语言处理与物联网技术中的边缘计算等方法提升对场景需求的精准分析与细化把控。同时，对海量数据实现价值分层，结合深度学习构建数据价值分层管理体系，完成从低质量数据到高价值数据的高效筛选与价值萃取。

其次，继续完善数据基础设施与自研工程，从底层基础层面提升数据存储与计算能力。可利用尖端大模型突出的文本数据集标注能力与经济化效益，持续打造契合国产人工智能发展脉络的中文场景训练数据语料库。同时，配合数据质量评估体系，对专业数据、平台数据等内容以特征导向展开质量评测，避免数据偏差与“虚假多样性”等数据风险。

最后，人工智能等现代技术的积极运用是提升数据场景化加工能力的核心关键。例如，在医疗、金融等不适合进行数据自由流动的敏感场景，可利用人工智能转化为“伪数据”或匿名数据的形式，结合大模型生成合成数据或进行数据增强。同时，运用隐私计算、联邦学习与分布式训练等技术方法，在保障数据安全的基础上提升数据的场景化加工与供给能力，在回应场景化数据需求的同时平衡数据“可用与可控”之间的矛盾。此过程中，须将人工智能人才培育作为重点政策并细化培养方案，积极运用专项基金等激励措施，带动专业化人才发展以实现场景需求下数据集建设与供给能力的跃升。

其二，在场景面向下推进公共数据、平台数据的可信开放与充分共享。一方面，我国阿里、腾讯、百度、字节跳动等超大型平台企业对海量互联网数据形成支配，但受到个人信息保护与平台成本的限制，简单要求共享数据的思路亦不可行。为实现少数平台数据垄断向全行业数据良性共享的过渡，首先，可借鉴欧盟与美国等西方数据治理相关立法中的“守门人”制度，将满足特定条件的大型平台企业规定为“守门人”并要求其承担特定的法律义务。进一步细化“守门人”制度的限制，明确“守门人”将非个人数据共享给第三方的要求及其具体规定。

其次，亦可借鉴知识产权制度中的保护时间设计，根据不同场景下的数据价值变化曲线要求，引导或鼓励平台在一段时间后共享具有公共属性的数据以便其他主体进行利用。

最后，进一步完善数据交易制度、数据产权制度与数据收益分配制

度。协调场内交易与场外交易等相异场景需求下的具体措施，逐步形成面向数据资源化、资产化、资本化等不同层面，兼顾薪资分配、效益分配和股权分配等多种分配形式的分配机制与交易生态。

另一方面，推动公共数据标准化、共享机制和安全保障体系建设，兼顾政府、企业和个人等主体对公共数据开放的需求与利益诉求，构建公共数据开放与共享的深度协同机制并提升跨部门协调能力与政策执行力。

其三，聚焦重点场景，赋能多样化场景，推进政府主导与市场平衡下的前沿数据利用。针对场景面向下公共数据与高价值数据的利用，政府不宜对市场“无形之手”的灵活调控效果进行过度干预，但也应对数据开放、流通与利用过程的合规性与安全性承担主要监管职责。可以采用“两步走”的方式推动从重点场景到多样化场景中良性秩序的形成。

第一步，现阶段暂时摒弃“不加区分、一并推进”的做法，明确数据集供给的划定应用场景。例如，《“数据要素×”三年行动计划（2024—2026年）》

中明确指出，要“聚焦重点行业和领域，挖掘典型数据要素应用场景”。通过借鉴域外取得良好效果的重点场景，可在地方层面进一步将高质量场景数据集建设明确在普惠金融、医疗保险、物业服务等相对有限的细分板块。在此基础上，完善重点应用场景下的数据开放细则，争取实现相应领域数据的有需必应。

第二步，在着眼核心场景的同时持续推进更大范围场景的赋能效应。

一是建立开放清单动态调整机制，同步维持开放数据的实时更新与日常性维护工作，确保实时分析能力的精准高效与相应救济途径的畅通。

二是以市场为主体推动多样化场景数据的衍生发展，先通过提高相似场景中的数据泛化表现以提升场景数据集的通用性，避免智能模型等末端应用过拟合于某些特定子场景。

三是针对元宇宙、数字孪生、智能决策等数字时代的新型场景进行数据需求的前瞻性剖析，并对智能监控等数字城市治

理场景中出现的的前沿问题作出精准把握，提前布局以保障新兴领域预见性政策的及时跟进。

其四，优化现有规范，明确数据架构与技术配套的场景化标准。

一方面，持续完善顶层制度设计以发挥其对场景数据流通的保障与激励机能。可参考美国《信息自由法》《联邦数据战略》与欧盟《通用数据管理条例》《数据治理法案》等规范中关联的可借鉴规则，构建公共部门信息复用与数据开放的场景化框架并推动本土化创新，鼓励行业导向的战略性数据集建设与数据开放生态。同时，在跨境电商、跨国物流、国际金融等场景中，跟进关注《中欧全面投资协定》(CAI)《全面与进步跨太平洋伙伴关系协定》(CPTPP)《区域全面经济伙伴关系协定》(RCEP)等国际经贸协议中的数据跨境流动规则，通过跨境场景的数据协同构建我国数据要素的独特场景优势，合理分配国家的“剩余控制权”，保留更多的行业自律与跨行业协同空间。

另一方面，持续完善场景面向的数据架构与技术标准。

一是推动差异场景下技术架构的统一化进程，促进高质量数据的跨场景互通，推动全国数据要素市场一体化。同时，亦不可忽视场景间的差异性，避免盲目追求共性而损害数据质量，跟进补充对标特定场景的数据标准，特别关注大模型背景下多模态数据的可识别性与技术架构互通等问题。

二是在技术标准制定进程中以前瞻视角关注人工智能等前沿技术的新型应用场景与产业模态。例如，明确不同场景下人工智能合成数据的质量标准；细化无监督机器学习中场景数据抓取的合规技术标准；厘清大模型生成内容领域数据集提供者、模型开发者与服务使用者之间的数据义务等。

三是以场景为牵引，建立强制性的“数据法规”与自愿性的“数据标准”相结合的新型标准化体制。特别是在医疗、科研等专业细分场景的数据标准中，运用弹性较高的技术标准与伦理标准发挥“软法先行”的正向效应，动态探索各数据场景的适配规则。⁸

(文章来源：中国城市经济发展微信公众号)