

编者按

今年政府工作报告首次提出“打造智能经济新形态”，并指出要深化数据资源开发利用，健全数据要素基础制度，建设高质量数据集。为探索工业数据“采”“集”“用”有效路径，工业和信息化部近日印发通知，启动工业数据筑基行动，开展面向人工智能赋能的高质量行业数据集建设先行先试，业界反应热烈。安徽省工业和信息化厅二级巡视员、安徽省新经济联合会会长潘峰特为本报撰写文章，就高质量数据集建设路径展开讨论。

“三二一”路径推动高质量数据集建设

安徽省工业和信息化厅二级巡视员、安徽省新经济联合会会长 潘峰

在“人工智能+”行动深入推进的时代背景下，高质量数据集已成为驱动产业智能化转型、培育新质生产力的核心要素。其价值不仅在于数据本身的精准性与标准化，更在于通过集群化效应实现从单点突破到面状发展的跨越式赋能。如何构建可持续、可推广、可落地的高质量数据集发展体系，推动高质量数据集建设收获“规模红利”，并向“质量红利”“生态红利”跃迁？笔者在此提出“三二一”思路，希望能为这项创新性工作做些抛砖引玉的尝试。

相关链接

工业数据筑基行动 2026年目标

到2026年年底

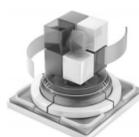
培育一批行业数据合作联合体，建设重点行业数据可信互联平台



汇聚一批行业数据资源，攻关一批数据关键技术



研制一批工业数据标准，打造一批高质量、标准化、可流通的行业数据集



赋能一批行业大模型、工业智能体等应用落地



总结形成工业数据高效采集处理、可信流通汇集、深度融合应用的有效路径、创新机制和经验模式，赋能行业提质增效，探索工业领域高质量数据集建设和数据开发利用模式。

问题导向是高质量数据集的立身之本。数据集的成功，关键在于精准切入全产业链发展中的痛点，有针对性地解决生产决策、供需对接、研发制造过程中的难点问题。合肥的杰士杰公司是一家生产改性工程塑料和高分子复合材料的公司，在企业发展过程中，一个难题就是，客户越来越多，需求品种已经超过了5000种，每一个品种的配方都要千百次地研发尝试，科研成本效率亟待提升。后来公司找到芜湖智唐科技公司，该公司直面这个棘手问题，从相关数据搜集、清洗、统一标准、标注等开始，建立起高质量数据集，然后用高质量数据集“喂”出相应的工业模型，最终打造出工业智能体，一举解决了这个问题。现在，相应的产品配方都是AI自动生成的了，研发效率提升80%以上，大大节约了科研成本。杰士杰公司对此就感到惊喜，立刻又着手和智唐公司合作解决其他棘手问题，目前已打造出三个智能体，全面提升了企业效率。这个案例让我们充分认识到，只有聚焦企业的难点、痛点问题，所构建的高质量数据集才能体现出价值，使之真正成为

当前高质量数据集建设面临的最大难题就是共享与隐私之间形成的矛盾。龙头企业走出了高质量数据集建设之路，但是出于对企业隐私的顾虑，多数企业不愿意把自己的数据集拿出来共享。怎么解决？笔者提出可解耦与不可解耦两条路径并行发展之策，以期既兼顾中小企业的共享需求，又尊重龙头企业的隐私诉求，实现“普惠性”与“特殊性”的有机统一。

可解耦路径，主要是针对产业集群的共性需求，通过数据共享构建“普惠型”高质量数据集。在中小企业密集的产业群中，单一企业的数据规模小、覆盖面窄，难以形成高质量数据集，但众多企业的共性问题为数据共享提供了基础。通过行业联盟或第三方平台组织，引导企业贡献非涉密的共性数据，经过标准化治理，形成可剥离、可共享的数据集，能够有效降低中小企业的智能化转型成本。中国农业

高质量数据集建设涉及数据采集、治理、标注、应用等多个环节，跨越政府、企业、科研机构等多个主体，单靠任何一方都难以完成系统性建设，构建“政产学研金服用”联合体是破解协同难题的关键。联合体的核心价值在于整合各方优势，形成“数据资源+技术研发+场景应用+政策保障”的闭环体系。政府部门发挥政策引导与资源协调作用，通过出台扶持政策、制定行业标准、搭建公共服务平台等方式营造良好环境。研究机构则聚焦核心技术攻关，突破智能标注、数据安全流通、质量评估等关键技术，为数据集建设提供方法论与工具支撑。制造企业、数字化企业、软件企业，包括金融单位、相关服务单位等作为关联主体，既

所谓“三”，是指坚持问题导向、价值导向、市场导向这三个导向，以期锚定高质量数据集解决实际问题、创造核心价值、实现持续运转的生命根基。

坚持三个导向

为企业降本增效的“利器”，这也是人工智能赋能新型工业化的核心逻辑。问题导向下的数据集建设，拒绝“为建而建”的形式主义，坚持“有用才建”的务实原则，让数据真正成为破解产业发展瓶颈的“金钥匙”。

价值导向是高质量数据集的推广之要。我们今天有声有色地开展高质量数据集建设，是为了让更多的企业从中受益，最大程度地摊开价值之饼。如何做到呢？就是要从成功的个性案例中萃取出共性的东西来，然后用这个共性的东西，为更多的企业送去福音。第一步，基于

个性化的共性萃取。笔者在此提出“因式分解+取最大公约数+共性模型化处理”的方法。“因式分解”，就是把解决企业问题的数据集细分成若干单元因子。“取最大公约数”，就是把解决同一类问题（不同案例）的相同单元因子取出来。“共性模型化处理”，就是把同一类问题带有共性的单元因子处理成模型单元。第二步，借助个性化模型单元，结合企业具体情况，加以微调处理，形成众多不同企业的个性化解决方案。这样，就可以实现从单点应用到批量赋能的辐射效应，实现高质量数据集建设的价值最大化，实现解决一类问题，带动一片产业。合肥的

零壹数智公司正在这方面进行积极探索，并取得一定的成效，在为阳光电源、雅迪这样的大公司建设高质量数据集的过程中，已经形成若干共性的模型单元，这让它们有条件为更多的企业高效地送去AI赋能服务。

市场导向是高质量数据集的续航之基，旨在通过市场化运作逻辑实现可持续发展。数据采集、治理、标注、更新等数据全生命周期环节均需持续投入，单纯依靠政策扶持或公益投入难以长久。市场导向下的高质量数据集建设，不仅要考虑技术可行性，更要设计清晰的价值变现路径，通过数据产品、技术服务、授权使用等多元模式，让数据要素在流通中实现价值增值，为数据集的持续迭代提供不竭动力。杰士杰公司和智唐公司就有意把他们基于高质量数据集建设起来的三个智能体剥离出来，成立新的AI赋能公司，以一种新的业态开展普适性的服务。在这项崭新的数字技术实践过程中，我们希望能有更多的新模式、新业态与之匹配。

所谓“二”，是指“可解耦”和“不可解耦”两条路径，以期破解高质量数据集的推广难题。

尝试两条路径

科学院构建的企业典型作业场景多模态数据集，正是通过“申请-授权”的共享机制，整合多家生产单位的作业数据，支撑AI算法研发与智能装备创制，帮助相关企业节省劳动动力10%以上。这种路径的核心价值在于“聚沙成塔”，让中小企业无须单独投入高昂成本，即可享受高质量数据集的赋能，同时通过数据共享进一步丰富数据集的多样性与完整性，形成“共享-赋能-共赢”的

良性生态。我们今天已经建成许多产业集群，完全可以针对不同细分行业的共性生产问题，构建起跨企业的可解耦数据集，推动整个产业集群的智能化水平提升。

不可解耦路径，立足龙头企业的隐私保护与生态引领，通过“数据集+大模型+生态”模式实现“辐射型”赋能。龙头企业往往拥有海量高价值数据，但这些数据涉及商业秘密或核心技术，难以直接共享。对此，可以鼓励

龙头企业依托自身数据资源构建私有高质量数据集，训练行业专用大模型与智能体，通过生态合作的方式，带动上下游企业发展。西门子与微软联合推出的工业元宇宙计划，正是依托西门子的工业全生命周期数据构建专有数据集，通过大模型赋能生态合作伙伴的生产优化与创新研发。国内医疗领域的龙头医院联合科研机构构建的专病病例库，虽未直接开放原始数据，但通过模型输出诊断辅助方案、科研分析结果等方式，赋能基层医疗机构与相关企业，实现了“数据不流出、价值流出去”的效果。这种路径既保障了龙头企业的数据安全，又发挥了其技术与资源优势，通过带动100家、200家生态企业共同发展，形成产业生态的“领头雁”效应，同样是高质量数据集集群化赋能的重要实现方式。

所谓“一”，是指打造一个联合体，以期凝聚起高质量数据集的建设合力。

打造一个联合体

是数据集的使用者，也是数据资源的提供者与应用场景的贡献者，从各自角度为数据集建设提供支撑保障。在联合体机制下，各方通过联合攻关、成果共享、风险共担，能够有效破解“数据孤岛”、技术壁垒、标准不一等突出问题。例如，针对工

业高质量数据集建设中存在的标准缺失问题，可由政府主导、科研机构牵头、企业参与，共同制定数据采集、标注、质量评估等系列标准；针对数据安全流通难题，可通过联合体共同研发隐私计算、区块链等技术方案，在保障安全的前提下，促进数据要素流动。

最终，依托联合体构建的高质量数据集，能够形成“数据集-大模型-工业智能体”的技术链条，针对具体产业问题实现精准赋能，让数据要素的价值在协同创新中充分释放。

高质量数据集的建设不是孤立的技术工程，而是关乎产业升级、生态协同、可持续发展的系统工程。“三个导向”明确建设的价值逻辑，确保数据集“有用、能用、好用”；“两条路径”破解共享与隐私的矛盾，实现不同主体的差异化赋能；“一个联合体”凝聚了多方合力，为建设工作提供了组织保障。三者相互支撑、有机统一，可以共同构成高质量数据集从建设到应用、从单点到集群、从短期突破到长期发展的完整路径。

奋力谱写新型工业化发展新篇章