



农业高质量数据集评价框架与全生命周期治理路径

谢能付^{1,2,3}, 王晓丽^{1,2,3*}, 王馨^{1,2,3*}, 孟智博^{1,2,3}, 满芮^{1,2,3}, 赵忠贺^{1,2,3}, 吴昊^{1,2,3}, 吕前杰^{1,2,3}

1.中国农业科学院农业信息研究所,北京 100081;2.国家农业科学数据中心,北京 100081;3.农业农村长期因子综合观测数据汇交总枢纽,北京 100081

摘要: 农业领域高质量数据集是支撑农业智能模型研发、数字农业应用与科学决策的关键数据基础。现有农业数据集多服务于特定模型训练和局部场景应用,在任务边界、样本代表性、专业标注、跨源互操作、持续维护和 AI-ready 表达等方面仍存在不足。围绕农业数据强时空异质性、场景依赖性和多源耦合性的特点,本文构建“基础质量层—农业场景适配层—治理复用层”的层级化评价框架,将准确性、完整性、一致性、代表性、任务适配性、跨域泛化性、可追溯性、互操作性、可复用性和合规安全等维度转化为质量证据、量化指标和度量方法。在此基础上,本文进一步梳理农业视觉图像、遥感与地理空间、文本知识和多模态感知数据集的建设现状,揭示农业数据质量问题沿“上游根源—农业特异性成因—质量缺口—应用风险”的传导机制。面向农业大模型和多模态智能应用,本文提出农业 AI-ready 数据结构应从传统“样本—标签”扩展为“指令—输入—输出—证据—约束”,并强化时序动态对齐、多源语义嵌入、知识/语义对齐和幻觉抑制。最后,本文提出农业高质量数据集全生命周期数据质量治理与逆向溯源框架,强调通过任务定义、分层采样、专家复核、质量评价、版本管理、用户反馈和分级开放,实现数据集从一次性资源建设向持续数据质量治理转变。

关键词: 农业数据集; 数据质量; AI-ready; 质量评价; 全生命周期数据质量治理; 农业人工智能; 开放共享

1 引言

近年来,以物联网感知、农业遥感、无人机巡田、智能农机和农业人工智能等为代表的数字农业技术快速发展。农业数据在生产记录、资源监测、病虫害识别和经营决策支撑等方面发挥着重要作用,推动了农业科研、生产及管理领域信息协作方式的变革^[1-4]。智慧农业大数据研究表明,农业数据已经贯穿传感器监测与云计算平台,并延伸至农产品供应链及农业经营管理等诸多环节^[1];与此同时,机器学习在产量预测与病虫害识别、土壤及杂草管理,以及农产品品质评价等方面的应用也不断拓展^[2]。然而,随着农业数据规模的扩大和农业智能模型能力的提升,数据集质量问题逐渐引起关注。

区别于通用互联网数据或传统统计数据相对稳定的数据形态,农业数据具有明显的时空异质性、场景依赖性和多源耦合性,其数据质量治理难度较

大。一方面,未经规范采集和质量控制的农业数据集可能存在样本偏倚、标签错误、字段缺失及噪声干扰等问题,进而影响模型训练和农业决策结果;另一方面,经过初步清洗和整理的数据集仍可能存在任务边界不清、场景覆盖不足、跨地区泛化能力弱以及长期复用困难等问题,从而在实际应用中影响生产判断。农业开放数据研究显示,不同平台和数据源之间在数据质量与元数据完整性,以及访问机制和共享条件上仍存在明显差异^[5]。以数据为中心的人工智能研究也指出,模型可靠性不只由算法结构决定,也受到数据采集、清洗、标注、增强、评估和维护等环节的共同影响^[6]。

与此同时,面向农业人工智能的模型训练、基准评测和跨场景应用需求持续增长,对现有农业数据集建设提出了更高要求。研究者发现,机器学习数据集不能只追求规模和性能指标,还应关注采集逻辑与样本代表性,兼顾数据许可及复用机制,并充分考虑后

收稿日期: 2026-07-20; 录用日期: 2024-09-15

基金项目: 中央级公益性科研院所基本科研业务费专项(JBYW-AII-2026-46),中央级公益性科研院所基本科研业务费专项(JBYW-AII-2026-47),2026 国务院国资委“揭榜挂帅”项目——“面向未来产业的国有企业数据资产产权界定与价值实现机制研究”(2052702826740711426)。

联系方式: 谢能付, E-mail:xienengfu@caas.cn。通信作者王馨, E-mail:821012530568@caas.cn; 王晓丽, E-mail:wangxiaoli@caas.cn。

续维护与潜在影响^[7]。农业病虫害图像数据集研究进一步表明,现有数据集中普遍存在类间样本不均衡、选择偏差、目标尺度变化、图像质量差异、标注困难和数据集可用性不足等问题,这些问题会直接影响识别算法的准确性和鲁棒性^[8]。在农业遥感与资源环境、文本知识及多模态生产等诸多数据类型中,类似风险还表现为时空基准(坐标系统与时间尺度)不匹配、字段语义分歧,以及质量记录缺失与跨平台互操作受限。不断扩展的农业智能应用场景,给现有数据集建设方式带来了挑战。

农业数据集的数据质量治理已成为关系农业智能化应用可靠性、农业科技创新和粮食安全治理的重要议题。在此背景下,农业高质量数据集相关研究发展迅速且分支繁多。除了显性的图像、遥感、传感器和文本等数据资源建设,研究者普遍将元数据规范、开放共享、数据权属、合规安全和持续维护等治理问题也纳入高质量数据集建设范畴。农业大数据标准体系研究指出,农业大数据应用需要涵盖基础框架、采集处理到管理运营,以及共享服务与应用等层级的标准支撑,从而保障多设备采集结果的可比性,同时提升多源数据的兼容性与多类分析系统的集成性^[9]。但从现有实践看,农业高质量数据集仍面临资源分散与标准各异的困境,且伴随着质量证据欠缺及可持续维护机制薄弱等现象,面向复杂农业场景的数据集建设仍处于不断完善的发展阶段。

数据驱动的治理范式依然是农业人工智能和数字农业应用的核心技术路径。农业智能模型和农业数据产品的有效性高度依赖于高质量、细粒度、可追溯的数据集。农业数据集的规模与质量,将直接决定农业智能模型泛化能力、数字农业平台服务能力和农业科学决策可信度的上限。针对这一现状,本文对农业领域高质量数据集建设进行系统性梳理。首先厘清农业高质量数据集的概念内涵并对其评价维度进行概述。进而,本文重点综述当前农业视觉与病虫害数据集,以及农业遥感与资源环境、农业文本知识与多模态生产管理等典型数据资源的建设现状。最后,本文分析农业数据集建设面临的主要问题,并从任务需求与采集设计、数据标准化治理与专家标注、质量评价、开放发布及持续维护等关键环节提出未来建设路径,以期对农业数据资源的统筹建设、开放共享与持续治理提供参考。

2 高质量农业数据集的内涵与评价维度

针对目前农业数据集由规模扩张转向数据质量治理的发展情况,本文从概念界定、评价维度、FAIR与AI-ready表达三个角度,对农业高质量数据集的内涵进行梳理。总体来看,高质量农业数据集并不是简单的大规模样本集合,而是能够支撑农业智能模型训练与生产诊断,以及资源环境监测和科学决策的数据资源体系。其核心问题在于,如何判断一个数据集不仅“能用”,而且“可信、可解释、可复用、可持续”。

为避免术语混用,本文将“数据质量治理”作为贯穿农业数据集建设全过程的总称,强调从任务定义、采集设计、清洗标注、质量评价到开放维护的持续治理过程;“质量评价”主要指基于指标体系和度量方法的数据质量测度;“质量控制”主要指采集、标注、评测和发布等阶段的具体控制节点;“数据治理”则主要用于描述数据标准、权属、开放共享和合规安全等制度性安排。除引用TDQM理论时保留“质量管理”表述外,全文统一采用上述术语体系。

2.1 概念界定

高质量农业数据集是指围绕特定农业任务构建,经过规范采集、专业标注、质量评价、标准化描述、合规发布和持续维护,能够支撑可靠分析、模型训练、跨场景复用和应用反馈的数据资源体系。其组成涵盖图像、遥感影像、传感器记录、土壤气象数据、农业文本和经营管理数据等数据本体,也包含标签体系、元数据、采集方案、标注规范、质量报告、许可协议、版本记录和维护机制等配套资料。

当前农业数据集建设中容易出现一种误区,即将“数据规模大”直接等同于“数据质量高”。但数据质量研究表明,数据质量具有明显的使用情境属性,需要从使用者需求和任务目标出发加以判断,而不能只以准确性单一指标衡量^[10-11]。在大数据环境下,数据来源异构、规模扩大、更新频繁和结构复杂还会进一步放大质量评估难度^[12]。因此,农业数据集的高质量并不是建成之后的静态结果,而是贯穿采集与清洗、标注及存储,直至发布和复用全过程的动态治理能力。

与通用数据集相比,农业数据集具有更强的场景依赖性。作物生长与病虫害发生、畜禽行为及土壤水分,以及农田边界和农产品品质等对象,都会受到区域跨度、季节变换、品种差异与生育期等自然属性影响,同时也受制于气象条件、栽培制度、管理措施及采集设备等外部因素。同一作物或同一病害在不同地区、不同年份和不同生育阶段可能呈现明显差异。农业开放数据研究显示,不同平台之间在元数据、质量控制和共享机制方面仍存在差异^[5];农业病虫害图像数据集研究也指出,类间不均衡与选择偏差,叠加尺度变化及可用性不足等问题,均会影响模型泛化能力^[8]。这意味着,高质量农业数据集必须同时回答“数据是否准确”和“数据适用于何种农业场景”两个问题。

面向农业人工智能和智慧农业应用,数据集还应具有明确的任务边界和评测边界。数据中心人工智能研究强调,模型性能不只由算法决定,也取决于数据采集与清洗、标注及增强,以及后续的验证和维护等环节^[6]。机器学习数据集研究也指出,数据集开发应关注采集逻辑与代表性,并兼顾许可授权、复用性及潜在影响^[7]。因此,本文认为,高质量农业数据集的关键不是单纯扩大样本数量,而是在明确农业任务的前提下,将数据质量、场景适配、开放复用和持续治理统一起来。

2.2 农业高质量数据集的层级化评价框架

农业高质量数据集评价体系的构建需要具有明确的理论基础。数据质量研究认为,数据质量并不是单一的准确性问题,而是由内在质量、情境质量、

表示质量和可访问质量等多个维度共同构成的综合属性^[10]。其中,内在质量强调数据本身是否准确、可信和一致;情境质量强调数据是否符合特定使用任务和应用场景;表示质量强调数据结构、格式和语义表达是否清晰;可访问质量则强调数据是否能够被发现、获取和使用^[13]。TDQM(Total Data Quality Management)进一步将数据质量视为一个动态管理过程,强调通过“定义—测量—分析—改进”的循环机制实现数据质量的持续提升^[14-15]。这一思想表明,数据集质量评价不应停留在静态结果判断,而应贯穿任务定义、数据采集、清洗治理、标注评测、开放发布和持续维护全过程。由此,本文构建“基础质量层—农业场景适配层—治理复用层”的层级化评价框架,以刻画农业数据集从数据形成、任务应用到持续治理的质量要求。

在上述框架中,基础质量层具有前置性作用。准确性、完整性和一致性直接决定数据是否具备基本可信度,若原始观测、标签或字段结构存在明显错误,后续的场景适配和开放复用便缺乏可靠基础。农业场景适配层体现农业数据集的领域特征,重点评价数据能否覆盖真实农业生产中的时空差异、生物生长差异和管理措施差异。治理复用层则面向数据集的长期使用和持续更新,强调数据来源、处理过程、标准格式、许可协议和版本记录等治理条件。三类层级之间构成由“数据可信”到“场景可用”再到“长期可复用”的递进关系。

农业高质量数据集质量评价的总体逻辑如图1所示,具体评价维度、核验要点和主要依据见表1。

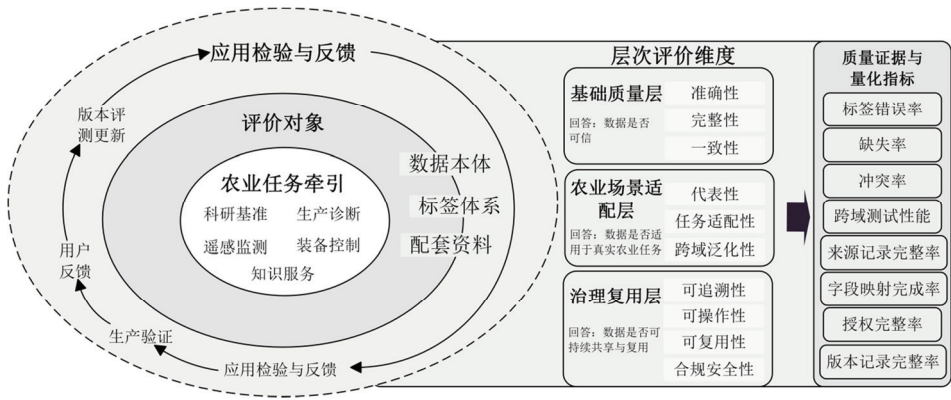


图1 农业高质量数据集层级化评价框架

Fig. 1 A hierarchical framework for evaluating high-quality agricultural datasets

表 1 农业高质量数据集层级评价指标体系：质量证据、量化指标与度量

Table 1 A hierarchical evaluation indicator system for high-quality agricultural datasets: quality evidence, quantitative indicators, and metrics

层级	评价维度	核心问题	质量证据	可量化指标与度量方法	短板指向
基础质量层	准确性	观测值和标签是否可靠	专家复核记录、地面调查结果、传感器校准说明	标签错误率=错误标签数/抽检样本数；专家一致率；传感器校准误差	干扰模型训练和评测
基础质量层	完整性	字段、样本和元数据是否充分	元数据清单、字段字典、缺失值处理记录	字段缺失率=缺失字段数/应填字段数；样本缺失率；元数据完备率	削弱复现和复用价值
基础质量层	一致性	空间、时间、语义和格式是否统一	坐标系统说明、类别字典、标注边界规则	格式冲突率；类别冲突率；坐标/单位一致率；模式冲突率	增加数据融合成本
农业场景适配层	代表性	样本是否覆盖真实农业差异	分层样本清单、跨区域划分、长尾样本说明	区域覆盖率；季节覆盖率；生育期覆盖率；长尾类别覆盖率；分布散度	造成跨场景泛化不足
农业场景适配层	任务适配性	是否匹配具体农业任务	任务说明书、标签体系说明、评测任务定义	任务标签匹配率；关键场景覆盖率；任务边界完整率	导致任务定义泛化
农业场景适配层	跨域泛化性	是否支持跨区域、跨年份、跨设备应用	跨域测试集、独立验证集、外部评测结果	跨域性能下降率=域内性能-域外性能；跨区域测试精度；跨年份测试精度	生产场景应用失效
治理复用层	可追溯性	数据来源和处理过程是否可审计	来源记录、处理流程、版本日志	来源记录完整率；处理流程记录率；版本记录完整率	降低数据可信度
治理复用层	互操作性	是否支持跨平台和多源融合	机器可读元数据、标准格式、字段映射表	字段映射完成率；标准格式覆盖率；机器可读元数据覆盖率	限制跨平台融合
治理复用层	可复用性	是否便于再次使用和复现实验	DOI、数据论文、基准划分、样例代码	文档完备率；样例代码可运行率；基准复现率	难以支撑长期研究
治理复用层	合规安全	是否具备开放边界和安全控制	脱敏规则、授权记录、用途限制说明	脱敏覆盖率；授权完整率；访问控制记录完整率；用途限制说明完备率	阻碍安全开放共享

图 1 表明，农业高质量数据集评价由基础质量层、农业场景适配层和治理复用层构成。其中，基础质量层强调数据本体可靠，农业场景适配层强调真实生产差异覆盖，治理复用层强调数据集的开放共享、持续维护和长期复用能力。

表 1 中，质量证据主要用于支撑指标计算和过程审计，可量化指标则用于衡量数据集在不同质量维度上的实际表现。需要强调的是，质量证据并不等同于评价指标。例如，分层样本清单只是判断代表性的证据来源，真正用于度量代表性的指标应包括区域覆盖率、季节覆盖率、生育期覆盖率、长尾类别覆盖率和样本分布散度；专家复核记录只是判断准确性的证据来源，具体评价时可进一步计算标签错误率和专家一致率；元数据清单也应进一步转化为字段缺失率、元数据完备率和关键农业过程信息覆盖率等指标。由此，农业高质量数据集评价可以从“证据清单式核查”进一步转向“证据支撑下的量化评价”。

为进一步增强评价框架的可操作性，本文引入层级加权思想。层级权重和指标权重可根据具体农业任

务、数据类型和应用风险进行设定，也可结合专家咨询、层次分析法或熵权法等方法确定^[16]。设农业数据集综合质量得分为 Q ，第 k 个质量层级的权重为 W_k ，该层级下第 j 个评价指标的权重为 W_{kj} ，标准化后的指标得分为 S_{kj} ，则有：

$$Q = \sum_{k=1}^K W_k \sum_{j=1}^{m_k} w_{kj} S_{kj}$$

其中， K 表示质量层级数量， m_k 表示第 k 个层级下的指标数量。对于覆盖率、完备率、一致率、复现率等正向指标，可采用正向标准化处理；对于缺失率、错误率、冲突率、跨域性能下降率等负向指标，可采用反向标准化处理，从而形成 0—1 之间的可比指标得分。层级权重和指标权重可根据具体农业任务、数据类型和应用风险进行设定，如表 2。

表 2 中的权重并非固定取值，而是用于说明不同农业任务对质量维度的侧重点存在差异。遥感作物分类与估产任务通常依赖多时相遥感影像、地面样本和气象环境数据，因此更强调时空覆盖、地面真值一致性和跨区域代表性。病虫害识别任务则更依赖专家标

签、病害阶段划分和田间复杂背景覆盖，因此准确性和农业场景适配性的权重相对较高。对于畜牧多模态监测和农业文本知识问答，模态配准、来源可信度、语义一致性和合规治理也会成为影响数据集质量的重要因素。

农业高质量数据集评价还需要考虑不同维度之间的权衡关系。第一，准确性与代表性之间存在一定张力。扩大区域、季节、生育期和复杂田间场景覆盖，有助于提升样本代表性，但同时会增加标注难度和标签噪声，因此需要通过专家复核、难例会审和分层抽检降低代表性扩展带来的准确性损失。第二，完整性与采集成本之间存在权衡。增加品种、生育期、气象、土壤和管理措施等元数据，有助于提升数据解释力和复用价值，但也会增加采集、清洗和维护成本，因此应根据任务目标确定最小必要元数据集。第三，开放复用与合规安全之间存在制约关系。农户经营数据、养殖场视频、高精度地块边界等数据具有一定敏感性，难以直接完全开放，需要通过脱敏开放、授权访问和分级共享机制平衡数据复用价值与安全治理要求。第四，互操作性与农业语义精细表达之间也需要协调。统一字段和格式有利于跨平台使用，但农业对象、病虫害类型、生育期和农事制度具有较强专业性，因此需要在通用标准和农业领域语义之间建立映射关系。

表 2 不同农业任务下质量评价权重侧重

Table 2 Weighting priorities in quality evaluation across different agricultural tasks

农业任务	基础质量层	农业场景适配层	治理复用层	权重侧重说明
遥感作物分类与估产	0.30	0.45	0.25	更重视跨区域、跨年份、时序完整性、地面真值一致性和空间代表性
病虫害识别	0.40	0.40	0.20	更重视标签准确性、病害阶段覆盖、田间复杂背景和长尾类别样本
畜牧多模态监测	0.35	0.35	0.30	更重视传感器校准、模态配准、个体追踪、连续记录和隐私合规
农业文本知识与问答	0.35	0.30	0.35	更重视术语一致性、来源可信度、语义对齐、可追溯性和开放复用

2.3 面向 AI-ready 的农业数据集结构化表达

当前农业数据集建设已开始重视开放共享，但从“可共享、可复用”走向“可训练、可推理、可评测”仍存在不足。FAIR 原则强调科学数据应具备可发现、可访问、可互操作和可复用能力，其核心作用在于通过持久标识符、标准元数据、规范协议、语义互操作和清晰许可，提升数据共享与跨主体复用能力^[17]。农业 FAIR 数据研究表明，FAIR 原则有助于农业数据开放共享，但农业领域仍存在数据分散、元数据不足和应用价值转化不充分等问题^[18]。

需要指出的是，FAIR 原则主要面向数据共享和开放复用，而 AI-ready 表达更强调数据能否被模型直接读取、理解、训练和评测。因此，FAIR 是农业数据集开放复用的重要基础，但并不是 AI-ready 的充分条件。面向农业人工智能应用，高质量数据集还需要在时空对齐、多源融合、语义表达、任务标签和模型评测等方面形成结构化表达。

数据集文档化是连接 FAIR 原则与 AI-ready 表达的重要抓手。Datasheets for Datasets 提出，数据集应系统披露建设目的与数据组成，涵盖采集过程及预处理方法，并说明分发方式、维护安排与伦理限制^[19]。

Data Cards 进一步强调，面向负责任人工智能的数据集说明应覆盖数据来源与开发意图、标注方式及训练评测用途，并明确适用边界和风险信息^[20]。对于农业数据集而言，病虫害图像数据需要说明专家判读规则和症状阶段，遥感样本需要说明地面真值来源和时间窗口，多模态畜牧或果蔬数据需要说明设备参数与模态配准，以及光照条件及质量筛选规则。通过文档化，数据集建设中的隐性决策才能转化为可审查、可复现、可改进的信息。

随着农业遥感、农机视觉和多模态感知任务增长，AI-ready 表达也成为农业数据集建设的重要方向。OGC TrainingDML-AI 标准面向地理空间机器学习训练数据，提供涵盖训练、验证及测试数据的概念模型，并将分类体系、地面真值、数据准备过程、来源质量和机器学习任务元数据纳入标准描述^[21]。这为农业遥感作物分类、地块边界识别和灾害监测等任务提供了重要参考。

但农业 AI-ready 数据结构不能停留在统一元数据和字段编码层面，还需要进一步解决三类关键问题。

第一，农业时序数据需要动态对齐。遥感时序、气象观测、土壤墒情、农事操作和病虫害发生记录的时间粒度并不一致，应通过地块 ID、时间窗口和生育期标签进行关联。

第二，多源异构数据需要语义嵌入表示。图像、遥感、传感器、文本和管理记录应通过农业本体、知识图谱或标准术语表，将作物、品种、生育期、病虫害和管理措施等概念转化为可计算的语义标识。同时，多源农业数据融合还需要数据协调机制，对术语规范、字段编码、单位、时间尺度、空间坐标和质量规则进行统一或映射^[22]。

第三，面向农业大模型的数据集需要构造指令微调样本。传统“样本—标签”结构难以支撑农业问答、病虫害诊断、遥感解译和管理建议生成，应进一步构建“指令—输入—输出—证据—约束”的数据结构，并记录答案来源、专家审核和适用边界。

第四，农业大模型数据集还需要强化知识/语义对齐和幻觉抑制。AgriLLM 等研究表明，真实农户问答数据和领域化咨询语料对农业大模型具有重要价值^[23]；

ShizishanGPT 将检索增强生成、农业知识图谱和工具调用结合起来，说明农业大模型需要与结构化知识和外部证据相结合^[24]；AgroGPT 面向农业视觉语言任务构建专家调优数据集 AgroInstruct，表明多模态农业大模型还需要图像、文本和专家知识之间的指令化对齐^[25]。因此，农业 AI-ready 数据结构不仅要包含传统样本和标签，还应记录指令模板、证据来源、专家审核、风险提示和适用边界，以降低模型在病虫害诊断、农药使用建议、遥感解译和生产管理决策中的幻觉风险^[26-27]。

因此，农业 AI-ready 数据集应由数据本体、农业过程级元数据、语义标识、任务标签、指令样本、证据来源、专家审核记录、质量凭证和版本信息共同构成。其目标不是简单实现数据开放，而是使数据能够被模型稳定读取、正确理解、跨场景调用和可靠评测。由此，农业数据集建设需要从“可共享、可复用”的 FAIR 导向，进一步发展为“可对齐、可融合、可训练、可推理、可评测、可追溯”的 AI-ready 导向。农业 AI-ready 数据结构的关键技术环节如表 3 所示。

表 3 农业 AI-ready 数据结构的关键技术环节

Table 3 Key technical aspects of ai-ready data structures in agriculture

技术环节	主要问题	表达机制
时序动态对齐	遥感、气象、土壤和农事记录时间粒度不一致	地块 ID、时间窗口、生育期标签
多源异构融合	图像、遥感、传感器和文本数据格式差异大	跨模态样本 ID、统一数据字典
语义嵌入表示	作物、病虫害、生育期等概念不统一	农业本体、知识图谱、标准术语表
指令微调数据	传统“样本—标签”结构难以支撑大模型应用	指令—输入—输出—证据—约束
知识/语义对齐	大模型回答易脱离农业专业知识和应用语境	知识图谱、RAG 检索、证据来源标注
幻觉抑制与审核	农技建议、病害诊断和用药建议存在高风险	专家审核、风险提示、适用边界说明
模型评测适配	随机划分难以检验真实泛化能力	跨区域、跨年份、跨生育期测试集

3 农业领域数据集建设现状：数据模态与业务场景

针对目前农业领域数据集的建设情况，本文从农业视觉与病虫害害、农业遥感与资源环境、农业文本知识与多模态生产管理三个角度进行现状分析。整体来看，农业数据集建设正在由单一数据汇集向任务化、场景化和可复用的数据资源体系演进。其服务对象既包括病虫害识别与杂草检测，以及作物表型测量及畜

禽健康监测等生产任务；也涵盖作物分类与种植结构制图，并延伸至资源环境评估、农业知识服务及政策分析等科研与管理场景^[5-8]。

与通用图像、文本或地理数据集相比，农业数据集更强调生物对象、生产环境和时空过程之间的耦合关系。同一标签往往会受到时空因素(如地区与季节)、农学特征(如品种、生育期及栽培制度)，以及气象条件与采集设备的共同影响。因此，当前农业数据集建设的核心问题，已经不只是能否提供足够样本，而是能否覆盖真实场景、支撑专业标注、提供完整元数

据,并形成开放共享和持续维护机制。

3.1 农业视觉图像数据集

农业视觉数据集是农业人工智能研究中发展较早、应用最活跃的数据类型。当前相关数据集主要覆盖植物病害识别、害虫识别、杂草分类、作物器官检测、表型测量和农田异常识别等任务。其中,PlantVillage 是植物病害图像数据集建设的代表性案例,其公开叶片图像推动了基于深度学习的植物病害诊断研究,也证明了大规模公开图像数据对移动端病害识别的促进作用^[28-29]。

但需注意,早期农业视觉数据多采集于相对规范的背景和光照条件下。当面临田间复杂背景与遮挡,或处于不同拍摄设备及品种条件下时,模型仍可能出现泛化不足。针对这一问题,PlantDoc、IP102 和 DeepWeeds 等数据集开始强调自然场景、长尾类别和原位采集。其中,害虫和杂草数据集进一步揭示了类间差异细微与类内变化较大,且伴随背景复杂及类别不均衡等问题^[30-32]。这表明,未来农业视觉数据集建设,是继续扩大图像数量,还是优先提升真实场景覆盖和专家标注质量,是值得持续关注的问题。

在作物表型和农田航空影像方面,数据集建设也在由分类任务转向检测、分割和基准评测。GWHD 面向小麦穗检测,采用多地区、高分辨率 RGB 图像和标注数据支撑小麦穗计数及表型分析方法开发^[33]。Agriculture-Vision 则以航拍农田影像支撑农田异常模式语义分割,体现了高空间分辨率、RGB 与近红外通道、地块尺度标注和农业模式识别之间的结合^[34]。国内数据集也在从单一病虫害分类向综合任务拓展,例如南京农业病虫害图像识别数据集包含虫害检测、病害检测、病害分类和水稻表型分割等子集,反映出国内农业视觉数据集正在面向多任务、多作物和可复用基准发展^[35]。

3.2 农业遥感与地理空间数据集

农业遥感、资源环境与地块数据集是支撑农业空间监测和宏观决策的重要基础。当前此类数据集主要服务于作物分类与种植结构识别,并涵盖地块边界提取、灾害监测,以及产量估计与资源环境评价。与近景图像相比,遥感数据集更强调时序完整性与空间分辨率,并关注地面真值、坐标系统,以及作物类别体系与跨区域可迁移性。

在遥感作物分类方面,CropHarvest 面向全球作物类型分类,关注卫星数据获取、处理和标注样本不足

等问题,也揭示了农业遥感模型在空间、语义及时间等多维度上面临域偏移与标签不均衡挑战^[36]。

Sen4AgriNet 基于 Sentinel-2 时间序列构建多年度、多国家作物分类与分割基准,体现了遥感训练样本从单时相影像向多时相、多区域和标准化标签体系发展的趋势^[37]。EuroCrops 通过汇聚和协调欧洲多国作物地块开放数据,说明大规模农业遥感数据集建设需要解决不同国家在数据来源与作物类别上的差异,以及文件格式与元数据之间的协调问题^[38]。

资源环境数据集为农业遥感和生产管理模型提供底层变量,但在农业数据集体系中常被忽视。土壤与气候、地形与水分,以及管理措施等数据,直接影响作物生长、病虫害发生和产量形成。SoilGrids 2.0 通过全球土壤属性空间预测和不确定性量化,为农业模型提供土壤有机碳与全氮,以及 pH 值和质地等基础变量^[39]。WorldClim 2 提供全球气候表面数据和生物气候变量,可用于农业区划、生态适宜性分析和作物模型输入^[40]。这类数据集的质量关注点不同于图像数据,更强调空间分辨率与时间范围,兼顾变量定义及单位一致性,并涉及缺失值处理与不确定性表达。未来应进一步加强地面调查样本与统计数据,以及气象土壤数据与遥感时序影像之间的关联标识,提升跨尺度融合和模型复用能力。

3.3 农业文本知识数据集

随着农业知识服务、智能问答、农技推广和病虫害辅助诊断需求增加,农业文本知识数据集逐渐成为农业数据集的重要类型。当前农业文本知识数据集主要服务于农业知识图谱构建、命名实体识别、关系抽取、信息检索和智能问答等任务,数据内容多围绕作物、病虫害、农药肥料、品种、症状、时期和防治措施等专业对象展开^[41]。AgCNER 作为中文农业病虫害命名实体识别数据集,围绕农业病虫害领域构建了大规模中文标注语料,说明中文农业自然语言处理任务需要面向专业术语体系、实体边界和专家审核机制建设数据集^[42]。

与图像和遥感数据相比,农业文本知识数据集的质量难点更集中于语义层面。术语同义、实体嵌套、区域性表达、知识更新和来源可信度都会影响数据集的可用性。未来农业文本知识数据集建设应同步记录语料来源、清洗规则、标注指南、实体类型定义和审核流程,并建立面向农业知识更新的持续维护机制。

大模型时代,农业文本知识数据集的建设重点正

在从传统实体识别、关系抽取和知识图谱构建，进一步扩展到指令微调、证据对齐和专家审核。AgriLLM 针对真实农户咨询问题开展农业问答自动化研究，体现了农业大模型对真实场景、问题驱动型语料的需求^[23]；ShizishanGPT 将农业知识图谱、检索增强生成和工具调用相结合，说明农业文本数据集不仅需要语料和标签，还需要可检索知识、证据来源和专业工具接口^[24]。因此，未来农业文本知识数据集应从“文本—标签”扩展为“问题—答案—证据—来源—专家审核—适用边界”的结构，以支撑农业大模型的知识/语义对齐和幻觉抑制。

3.4 多模态感知与生产管理数据集

多模态感知与生产管理数据集是近年来农业数据集建设的新趋势，主要面向畜牧监测、果蔬采收、农机装备感知、病害综合诊断和生产管理决策等复杂场景。与单一模态数据不同，这类数据通常融合可见光图像、红外热成像、传感器记录、环境变量、管理记录和实验检测结果，因而更强调多设备同步、模态配准、连续采集质量控制和生产过程可追溯。

已有数据集体现了这一发展方向。例如，奶牛个体头部多模态图像数据集将可见光图像与红外热成像图像配准，用于奶牛个体检测、健康监测和精准养殖研究，体现了跨模态配准、质量筛选和动物自然活动状态采集的重要性^[43]。柑橘黄龙病田间症状识别图像

数据集覆盖自然生态、白色背景板和黑色背景板等采集环境，并结合专家分类标注和部分 PCR 检测结果，为病害图像数据集的标签可靠性提供了有益范例^[44]。不同成熟度番茄果实多模态图像数据集则面向采摘机器人和果实品质识别，强调不同成熟阶段、光照条件、检测与分割标注之间的协同^[45]。

总体来看，多模态感知与生产管理数据集的建设难点并不只是数据类型更丰富，而在于不同模态之间的空间配准、时间同步、模态缺失处理、连续采集质量控制和多任务标签一致性。未来这类数据集的质量评价应从单一标签正确性扩展到模态互补性、任务适配性、过程元数据完整性和生产过程可追溯性。

在多模态方向，农业视觉语言模型对图像—文本—专家知识对齐提出了更高要求。AgroGPT 利用农业视觉数据构建 AgroInstruct 专家调优数据集，并面向农业多模态问答和专家式解释进行模型训练，说明农业多模态数据集已不再只是图像、红外或传感器数据的简单汇集，而需要进一步构造图像描述、诊断依据、专家解释和任务指令之间的对齐关系^[25]。因此，面向农业大模型的数据集应加强多模态样本 ID、图像文本描述、专家诊断证据和应用约束的统一组织。

为概括不同农业数据集类型与应用场景之间的关系，表 4 选取具有代表性的数据集案例，归纳其对农业数据集建设的启示和暴露出的质量问题。

表 4 不同数据模态下代表性农业数据集案例及其综述启示

Table 4 Representative agricultural dataset cases across different data modalities and their review insights

数据模态	代表性案例	主要业务场景	主要价值	质量问题
视觉图像	PlantVillage、PlantDoc	病害识别	推动病害诊断模型发展	实验环境与田间环境迁移落差
视觉图像	IP102、DeepWeeds	害虫识别、杂草分类	凸显自然场景和长尾类别价值	类别不均衡、尺度变化
遥感与地理空间	CropHarvest、Sen4AgriNet	作物分类、遥感时序建模	验证多区域时序样本对泛化的意义	地面真值不足、域偏移突出
地块与资源环境	EuroCrops、SoilGrids 2.0、WorldClim 2	地块开放数据、土壤气候建模	支撑跨机构复用与环境变量建模	类别体系、分辨率和元数据标准不统一
文本知识	AgCNER	农业实体识别、知识服务	拓展农业中文 NLP 数据基础	术语边界、实体嵌套和来源可信度
多模态感知	奶牛多模态、柑橘黄龙病、番茄多模态	畜牧监测、病害诊断、果蔬采收	支撑真实生产场景感知	模态配准、专家审核和过程追溯不足

4 农业高质量数据集建设面临的主要问题

针对目前农业高质量数据集建设的发展情况，本

文从需求边界、样本代表性、标注评测，以及标准互操作与开放维护等五个角度分析其主要问题。总体来看，现有农业数据集数量持续增加，数据类型也从图像与遥感，扩展至文本知识、多模态感知及生产管理记录。但“能下载”“能训练”并不等同于“高质量”

“可复用”。农业高质量数据集建设仍是一项系统工程，贯穿于任务定义与数据采集，涵盖清洗治理及专业标注，并延伸至质量评价、开放发布与持续维护等环节。

需要指出的是，需求边界不清、样本代表性不足、标注可靠性不高、元数据不完整和开放共享受限等问题并非农业领域独有，在医疗、工业制造、智能驾驶等数据集建设中同样存在。农业数据集的特殊性并不在于这些问题的名称不同，而在于这些问题的形成机制、影响路径和治理要求具有明显的农业领域特征。具体而言，农业数据质量问题往往由生物生长过程、物候阶段变化、气候区差异、农作制度、品种差异、田间管理措施和多源观测条件共同塑造。同一作物、病害、虫害或生产状态，在不同生育期、气候区、种植制度和管理条件下可能表现出不同的数据形态和标签含义。因此，农业数据集质量问题不能仅从一般数据治理角度理解，还需要从农业对象的生物属性、生产过程的周期性和农业场景的区域耦合性出发加以分析。上述问题从

上游设计缺陷传导到下游应用风险的形成机制如图 2 所示。以病虫害识别数据集为例，若数据集建设初期仅将任务界定为“病害识别”，但未进一步区分叶片分类、田间目标检测、病斑分割和病害程度评估等具体任务，就会造成任务边界不清。受此影响，采集阶段可能更倾向于收集实验室背景下的典型病斑图像，而忽略不同生育期、病害初发阶段、田间复杂背景、不同光照条件和不同品种下的样本覆盖。进一步地，由于同一病害在不同生育阶段可能呈现不同症状，同一症状也可能由病害、虫害、缺素、药害或气象胁迫引起，如果标注规范未纳入生育期、病害阶段和专家诊断依据，就容易出现标签语义漂移和标注一致性不足。上述缺陷会在质量层面表现为样本代表性不足、跨域泛化性下降和农业过程级元数据不完整，并最终导致模型在真实田间环境中出现误诊、漏诊和测试集表现与实际应用表现不一致等风险。因此，农业数据质量问题并不是孤立存在的，而是沿着“任务定义偏差—采集设计偏差—标注语义偏差—质量缺口—应用失效”的路径逐级传导。

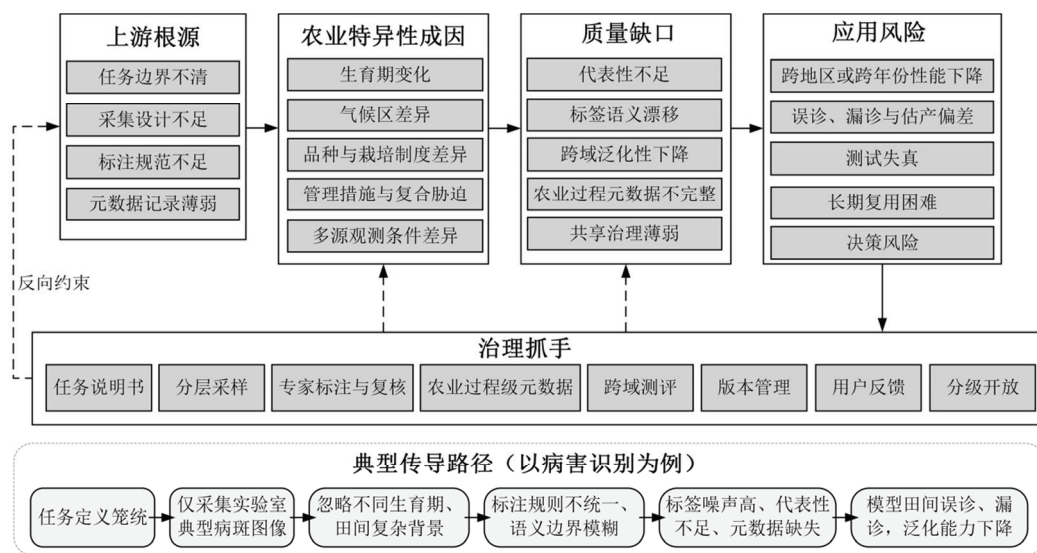


图 2 农业数据特异性质量问题的生成与传导机制

Fig. 2 Generation and transmission mechanisms of quality issues specific to agricultural data

图 2 表明，农业数据集质量风险并非由单一环节造成，而是由任务定义、采集设计、标注规范和元数据记录等上游缺陷，在生育期变化、气候区差异、品种与栽培制度差异等农业特异性因素作用

下，逐步传导为代表性不足、标签语义漂移、跨域泛化性下降等质量缺口，并最终影响模型评测可靠性和生产应用效果。通用数据问题与农业数据特异性质量挑战对照如表 5 所示。

表 5 通用数据问题与农业数据特异性质量挑战对照

Table 5 A comparison of general data issues and agricultural data-specific quality challenges

通用数据问题	农业特异性表现	深层成因	可能质量后果
需求边界不清	同一数据集难以同时服务病害分类、病斑分割、病害程度评估和防治决策	农业任务与作物生育期、管理目标、生产环节高度绑定	任务定义泛化，模型应用边界不清
样本代表性不足	样本集中于少数地区、季节、品种或理想拍摄条件	作物生长受气候区、栽培制度、品种和田间管理共同影响	跨区域、跨年份、跨生育期泛化能力下降
标签语义漂移	同一病害或长势标签在苗期、开花期、成熟期含义不同	生物生长过程导致表型连续变化，标签依赖物候阶段	标签边界不稳定，训练集与测试集语义不一致
域偏移	同一作物在不同气候区、不同种植制度下呈现不同影像或表型特征	气候、土壤、品种、轮作制度和管理措施共同改变数据分布	跨气候区、跨种植制度应用失效
标注可靠性不足	病虫害、作物长势、畜禽健康等标签需要专家结合症状、时期和背景判断	农业标签依赖植保、农学、畜牧、遥感等领域知识	普通标注员难以保证一致性，专家复核成本高
元数据不完整	缺少品种、生育期、气象条件、栽培制度、管理措施等农业上下文	农业数据解释高度依赖环境和管理背景	数据难以复用，模型误差难以解释
开放共享受限	农户经营数据、地块边界、养殖场视频等难以完全开放	涉及隐私、商业秘密、地理位置和生产安全	数据开放不足，长期维护和复用困难

表 5 表明，农业数据集面临的许多问题具有通用数据治理问题的外在形式，但其深层成因更多来自农业生产对象的生物属性、生产制度和区域环境耦合。因此，农业数据质量治理需要在通用数据规范基础上进一步嵌入物候、生育期、气候区、品种、栽培制度和管理措施等领域信息。

4.1 需求定义与使用边界不清

当前部分农业数据集以模型训练或论文实验为直接目标，建设初期对最终服务对象、应用场景和使用边界说明不足。农业数据集可能服务于科研基准、生产诊断与装备控制，或涉及遥感监测、经营管理及监管决策。不同任务对数据粒度、标签精度，以及时效性、错误容忍度与合规要求方面的需求并不相同。若数据集只说明数据类别和样本数量，却缺少对采集目的与目标用户的阐述，未界定适用区域及适用季节，且忽略设备条件与限制情形，使用者就难以判断其能否迁移到新的农业场景。

这一问题在农业领域较为突出。机器学习数据集研究指出，农业任务通常与作物生育阶段、生产环节和管理目标紧密绑定。例如，病虫害图像数据集若仅说明“病害识别”，仍不足以界定其任务边界，因为叶片分类、田间目标检测、病斑分割、病害程度评估和防治建议生成所需的样本粒度、标签体系和评测标准并不相同。遥感作物数据集同样需要区分作物类型识别、种植结构制图、长势监测、

产量估计和灾害损失评估等不同任务。若忽略生育期、区域气候、作物品种和管理措施等约束条件，数据集即使在原始研究任务中可用，也难以判断能否迁移到其他农业场景。

机器学习数据集研究指出，数据集开发不能只关注规模和性能指标，还应关注采集逻辑、代表性、许可、复用和潜在影响^[7]。Datasheets 和 Data Cards 也强调，数据集应披露建设目的与组成结构，详述采集过程及预处理方法，并说明分发维护机制、预期用途及风险限制^[19-20]。例如，实验室叶片图像训练出的病害识别模型，未必适用于田间复杂背景；单一区域地块样本训练出的作物分类模型，也可能难以适应异地种植制度和遥感时序差异。因此，需求定义不足是农业数据集从“研究可用”走向“生产可用”的首要障碍。

4.2 采集设计不足与样本代表性偏弱

农业数据天然具有强时空异质性。作物品种与生育期、栽培制度及土壤气候，叠加病虫害发生阶段与田间管理措施，以及传感器类型和采集人员习惯，均会影响数据分布。当前一些数据集的采集方案仍缺少分层设计，样本容易集中于少数地区及季节，或局限于特定品种与较理想拍摄条件，进而形成选择偏差。

与一般图像或文本数据不同，农业样本代表性不足不仅表现为样本数量不足，更表现为关键农业

差异维度覆盖不足。同一作物在不同生育阶段的冠层结构、叶色、纹理和遥感时序曲线可能明显不同；同一病害在初发期、扩展期和重症期的症状表现也存在差异；同一作物在单作、套作、轮作或复种制度下，其影像纹理、物候曲线和田间空间结构也会发生变化。因此，农业数据采集不能仅按照样本总量或类别数量衡量代表性，还应围绕区域、年份、季节、生育期、品种、栽培制度、管理条件和异常事件等维度进行分层采样。

已有研究对此给出了较为明确的证据。农业病虫害图像数据集综述指出，类间不均衡与选择偏差，以及尺度变化、图像质量差异及可用性不足，是影响数据集质量的重要问题^[8]。IP102、DeepWeeds等真实场景数据集也表明，害虫、杂草及病害对象往往存在类别长尾与类内差异大，且伴随背景复杂及目标尺度变化明显等特征^[31-32]。遥感数据集同样面临空间、语义和时间维度的域偏移，CropHarvest等全球作物分类数据集揭示了作物类别、地面样本和区域迁移之间的复杂关系^[36]。在国内实践中，区域性病虫害与棉田昆虫，以及畜牧场景及果蔬多模态数据集的出现，说明农业数据集正在向真实生产场景扩展，但低频病虫害与极端天气，加上异常管理措施及边缘地区样本仍然难以充分覆盖^[35,43-45]。代表性不足会造成模型在常见类别上表现较好，却在关键长尾场景中失效。

4.3 标注质量与基准评测可靠性不足

农业数据标注具有较高专业门槛。病虫害症状识别需要植保知识，作物表型和器官分割需要生物学与农学经验，遥感地块和作物类别标注需要地面调查、影像判读和时序分析能力，畜禽行为和健康状态标注也需要养殖专业知识。当前一些数据集虽然给出了标签文件，但统一标注指南与专家复核，以及难例处理规则及一致性评价仍不充分。

农业数据标注的特殊难点在于，标签含义常随生育阶段、病虫害发展阶段和生产环境发生变化，具有明显的阶段依赖性。以病害识别为例，同一病害在苗期、拔节期、抽穗期和成熟期可能呈现不同症状；同一症状也可能由病害、虫害、缺素、药害或气象胁迫等不同原因引起。此时，标签并不是单纯的视觉类别，而是需要结合生育期、发病部位、发生程度、环境条件和专家诊断共同确定。与智能驾驶中的天气、昼夜等外部域偏移相比，农业中的

生育期差异不仅改变观测外观，还可能改变标签含义和诊断标准。也就是说，智能驾驶中的天气和昼夜主要影响“看见的条件”，而农业中的生育期、病程阶段和管理措施还会影响“被识别对象本身的状态”以及“标签判定标准”。

标注质量不足会直接影响基准评测的可靠性。通用机器学习研究表明，测试集标签错误会干扰基准评测结果，并可能改变模型优劣判断^[46]。这一问题在农业数据中更值得警惕，因为农业对象常受遮挡与光照限制，面临相似症状及混合感染等复杂情况，加之作物重叠与地块边界模糊影响，人工标注不确定性较高。当前一些数据集虽然给出了训练集、验证集和测试集划分，但划分方式多以随机抽样为主，对跨地区与跨年份、跨设备及跨品种，乃至跨生育期泛化的检验不足。数据中心人工智能研究强调，数据采集与清洗、标注及增强，乃至最终评估都应纳入质量控制链条^[6]；农业病虫害数据集评价也提示，应从图像与标注质量，结合类别分布、可用性及模型验证等方面进行综合评价^[8]。因此，农业数据集不只需要提高训练样本质量，更需要保证测试集和基准任务的独立性、代表性和可复现性，并记录生育期、症状阶段、诊断依据和专家复核信息。

4.4 元数据、标准和互操作能力不足

农业数据集往往同时涉及遥感影像与无人机图像，涵盖近景图像、传感器记录及气象土壤数据，并融合农事管理记录、文本知识与空间矢量数据。不同来源数据在坐标系统与空间分辨率、时间粒度及计量单位方面存在差异，且在文件格式、字段命名，乃至作物类别与病虫害术语上不一致，会显著增加数据融合和复用成本。当前农业数据集建设中，元数据仍多停留在文件说明或简单字段表层面，对采集设备参数与地理位置、时间窗口及生育期描述不足，且缺乏针对传感器校准、标注人员信息，以及质检与版本变更记录的说明。

农业元数据缺失的影响比一般数据字段缺失更为复杂。对于农业数据集而言，品种、生育期、播种时间、采收时间、气象条件、土壤类型、施肥灌溉、病虫害防治措施和采集设备参数等信息，不只是数据描述字段，而是解释数据分布和模型误差的重要上下文。若遥感作物分类数据缺少物候期和种植制度信息，模型误差可能被误判为空间泛化能力不足；若病害图像数据缺少采集时期和发病阶段信息，标签差异可能被

误判为标注错误或模型识别能力不足。因此，农业数据集的元数据建设应从文件级描述扩展到“农业过程级描述”，将时空位置、物候阶段、品种制度、管理措施和采集条件纳入标准化记录。

从标准建设看，农业大数据标准体系研究指出，农业大数据应用需要标准保障多设备采集结果的可比性、多源数据的兼容性和多类分析系统的集成性，但农业大数据领域可直接指导实践的标准和规范仍显不足^[9]。从地理空间机器学习看，OGC TrainingDML-AI已将训练、验证、测试数据的任务、标签、地面真值、数据准备过程、来源、质量、版本和许可纳入概念模型，为农业遥感训练样本提供了可借鉴的标准化方向^[21]。数据协调研究也表明，多源数据复用需要统一或映射术语与字段，规范编码及单位，并建立一致的质量规则^[22]。如果缺少机器可读元数据和跨平台标准，农业数据集就难以进入更大范围的数据空间、模型评测体系和农业决策平台。

4.5 开放共享、合规安全和持续维护不足

开放共享是提升农业数据集复用价值的重要方式，但农业数据开放面临比通用数据更复杂的制度约束。农业开放数据综述表明，农业开放数据覆盖传感器、统计数据及卫星影像等多种类型，同时受开放数据不足与共享障碍制约，并受隐私关切、数据共享指南及FAIR符合度等因素综合影响^[5]。农业FAIR数据研究也提示，FAIR原则有助于农业数据共享和可持续绩效提升，但跨机构复用仍需解决数据分散、元数据不足和治理机制不完善等问题^[18]。

农业数据开放共享的特殊性还体现在数据权属和敏感性边界复杂。农业数据往往同时涉及农户个体、合作社、农业企业、地方管理部门和科研机构等多主体，数据内容可能包括农户经营记录、地块边界、养殖场视频、农机作业轨迹、投入品使用记录和产量收益信息。这类数据既具有科研和模型训练价值，也可能涉及个人隐私、商业秘密、地理空间敏感信息和生产安全。与一般公开图像或文本数据相比，农业数据开放不能简单采用完全公开方式，而应根据数据敏感程度采用完全开放、脱敏开放、授权访问和持续服务等分级共享机制。

持续维护不足是另一突出问题。农业数据集发布后仍需要处理失效链接与格式迁移，推进标签纠错、样本补充和版本更新，并兼顾用户反馈与引用追踪机制。由于农业生产具有季节性和年度循环特征，数据

集还需要持续补充新年份、新品种、新病虫害、新气候异常和新管理方式下的样本。研究数据仓储质量保障调查显示，数据质量保障具有多面性和非线性特征，需要审查流程、责任分工和资源投入支撑，且质量信息不足本身就是重要挑战^[47]。如果农业数据集缺少DOI与数据论文支撑，缺乏标准元数据、版本号、维护主体、变更日志和反馈渠道，就容易成为一次性项目成果，难以支撑长期复用和模型迭代。由此看，农业高质量数据集建设不能止于采集和发布，而应形成“开放共享、合规控制、质量反馈、持续更新”的治理机制。

5 农业高质量数据集未来建设路径

针对农业高质量数据集的发展现状和农业人工智能应用需求，本文从任务边界、数据代表性、标准互操作、标注评测、开放维护、面向大模型的数据结构以及全生命周期数据质量治理等方面提出未来建设路径。总体来看，农业数据集建设正处于由“数据汇集”向“数据质量治理”转变的阶段。未来建设的关键不在于单纯扩大数据规模，而在于明确农业任务边界的基础上，建立覆盖采集、标注、标准化治理、质量评价、发布共享、应用反馈和版本维护的全生命周期机制^[10-12,17-22]。农业高质量数据集全生命周期数据质量治理与逆向溯源框架如图3所示。

图3表明，农业高质量数据集建设不是单一机构或单一环节能够完成的线性流程，而是由科研机构、农业推广与生产主体、企业与平台、管理部门以及期刊/数据平台等主体共同参与的全生命周期数据质量治理过程。其中，科研机构主要参与任务定义、采集设计和质量评测，农业推广与生产主体主要提供真实场景样本、应用反馈和场景验证，企业与平台主要承担数据采集入库、技术支撑、发布共享和版本维护，管理部门主要负责标准规范、合规审查和开放分级，期刊/数据平台则支撑数据发布、DOI赋码和引用传播。

在治理流程上，任务定义、采集设计、数据采集与入库、标注治理、质量评测、发布共享和应用反馈与版本维护共同构成闭环。应用反馈并不是流程终点，而是反向修正的起点。当用户反馈或失败样本显示模型在特定区域、生育期、设备条件或生产场景下失效时，问题可反向回流至任务定义、采集设计、数据采集与入库、标注治理和质量评测等

环节,推动任务边界校正、样本补采、元数据补录、标注规范修订和测试集调整。右侧所列任务说明书、采集方案、样本覆盖清单、农业过程元数据、专家

复核记录、质量评测报告、用户反馈记录和版本日志等质量凭证,则为问题定位、责任追踪和版本修复提供依据。

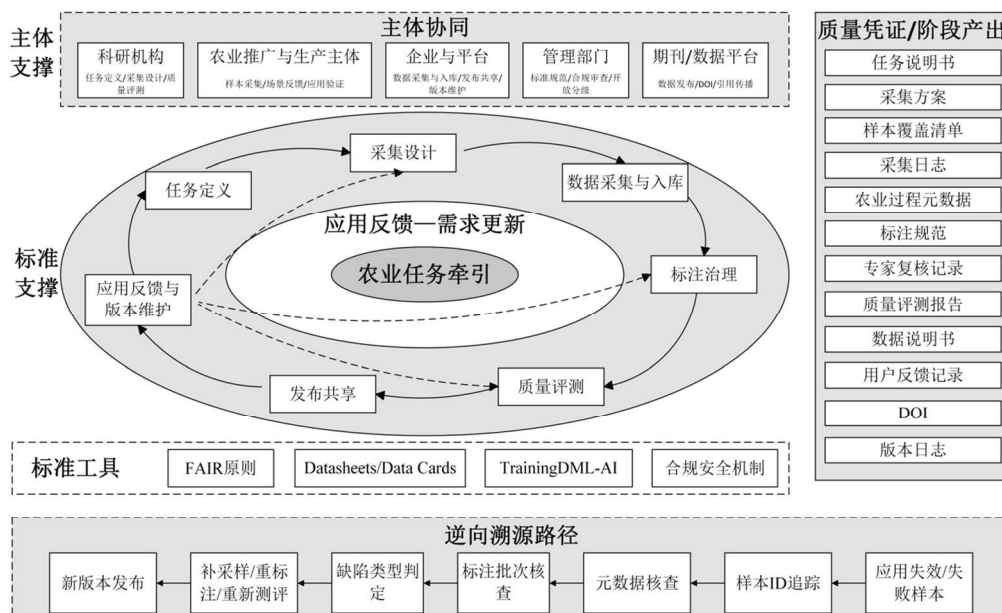


图3 农业高质量数据集全生命周期数据质量治理与逆向溯源框架

Fig. 3 A framework for full-lifecycle data quality governance and reverse traceability of high-quality agricultural datasets

5.1 任务边界清晰化

农业高质量数据集建设首先需要回答“为谁建、用于什么任务、在哪些条件下使用”的问题。当前部分农业数据集虽然能够支持模型训练,但其应用对象、适用范围和限制条件并不清晰,容易造成数据误用和模型迁移失败。未来数据集建设应在立项阶段形成任务说明书,明确服务对象与应用场景,界定空间及时间范围,并清晰地说明标签体系、评价指标、错误成本以及不适用情形^[19-20]。例如,病虫害识别数据集需要区分叶片分类与田间目标检测,以及病斑分割和病害程度评估;遥感作物数据集则需要区分像元分类与地块级识别,外加种植结构制图和灾害监测。只有先明确任务边界,才能避免以单一数据集承担过多功能,进而降低后续模型评测和生产应用中的风险^[6-7]。

5.2 数据代表性与规模性

尽管,农业人工智能仍高度依赖数据驱动,但农业数据集建设不能简单等同于样本规模扩张。农业对象受到区域与季节更替、品种及生育期差异、栽培制

度与管理措施,以及天气条件和设备类型的共同影响。如果采集过程过度依赖便利样本或理想场景,就会造成数据分布偏窄,进而削弱模型在真实农业场景中的泛化能力。已有病虫害、杂草和遥感数据集研究表明,类间不均衡与选择偏差,以及目标尺度变化和长尾样本不足是影响数据集质量的重要因素^[8,31-32,36]。未来农业数据集建设,是继续追求更大规模,还是在一定规模基础上提高样本覆盖和质量密度,是值得持续研究的问题。对此,采集方案应围绕区域与季节、品种及生育期、栽培制度以及异常事件进行分层设计,并重点补采低频病虫害与极端天气,或复杂背景及边缘地区样本。奶牛头部多模态、柑橘黄龙病田间图像和番茄成熟度多模态等数据集表明,真实场景中的多背景与多光照条件,结合多任务标注,能够显著提升数据集的应用解释力^[43-45]。

5.3 标准完备性与互操作问题

农业数据集的标准完备性是影响跨平台复用和多源融合的关键问题。当前农业数据往往同时包含遥感影像与无人机图像、近景图像,以及传感器记

录、气象土壤数据、农事管理记录与农业文本等类型。不同数据源在坐标系统与时间粒度、计量单位,以及字段命名、类别编码和文件格式上差异明显。如果缺少统一元数据和治理规范,后续模型训练、统计分析和开放共享都会面临较高成本。农业大数据标准体系研究指出,标准应覆盖基础、采集处理、管理、共享服务和应用等层级^[9]。未来数据集建设应强化机器可读元数据与数据字典,并保障采集规范、标注规范以及质量报告的同步发布。对于跨地区、跨平台和跨模态数据融合,还应先完成术语、字段、编码、单位和类别体系的映射,再进入模型训练或开放发布环节^[18]。

5.4 标注可靠性与质量评测

农业数据集的质量可靠性很大程度上取决于专业标注和基准评测。与通用图像或文本数据不同,农业数据标注常常需要农学与植保、遥感,以及畜牧和食品品质等专业知识。当前一些数据集虽然给出了标签和训练测试划分,但标注指南与专家复核机制,外加一致性统计及难例处理记录仍不充分。未来数据集建设应建立“标注员初标、领域专家复核,结合质检员抽检与争议样本会审”的流程,并保留复核意见和修改原因^{[8][19-20]}。对于测试集和基准评测集,还应采用更严格的专家复核和封闭管理,避免标签错误影响模型优劣判断^[46]。在质量评测方面,不应止于随机划分和单一准确率指标,还应设置跨区域与跨年份,或者跨设备、跨品种及跨生育期的测试集,以检验模型在真实场景下的稳定性^[6,8,33,36]。质量评测的目的也不应止于给出模型分数,还应反向发现数据缺口、标签错误和场景盲区。

5.5 开放共享与持续维护

开放共享是提升农业数据集复用价值的重要路径,但其前提是建立清晰的许可与引用,以及访问和维护机制。当前许多农业数据集在发布后仍面临链接失效与格式迁移、标签纠错及版本变更,以及用户反馈不足等问题。研究数据仓储质量保障经验表明,质量审查并不是一次性工作,而是需要结合责任分工、资源投入和用户反馈持续开展^[47]。未来农业数据集发布应配套 DOI 与标准元数据,提供数据说明文档及开放许可,明确引用方式与样例代码,并包含基准划分和基线结果,使数据能够被发现、理解、调用和复用^[17,19-20]。对于涉及农户经营、企业生产、养殖场视频、高精度地块位置和敏感地理信息的数

据,应采用脱敏处理与分级开放,结合授权访问、用途限制及审计记录,平衡共享价值与合规安全^[5]。WorldCereal 等开放作物制图系统表明,开放源码、可复现流程、季节性更新和质量置信信息正在成为农业数据产品复用的重要方向^[48]。

5.6 面向农业大模型的指令数据与证据对齐

面向农业大模型和多模态基础模型,未来农业数据集建设需要从传统“样本—标签”结构扩展到“指令—输入—输出—证据—约束”结构。其重点包括三方面:一是构建覆盖农技问答、病虫害诊断、遥感解译、生产管理和政策咨询等场景的高质量指令微调数据;二是建立农业知识图谱、标准术语、专家规则和外部文献之间的证据对齐机制,使模型回答能够追溯到可靠来源;三是引入专家审核、风险提示和不确定性表达机制,降低模型在农药使用、病害诊断和经营决策建议中的幻觉风险。农业大模型研究表明,真实农户问答、领域知识检索、工具调用和多模态专家调优数据正在成为农业 AI 系统的重要数据基础^[23-25,49]。

5.7 农业高质量数据集全生命周期数据质量治理与逆向溯源框架

农业高质量数据集建设不应停留在数据发布前的一次性质量检查,而应形成覆盖任务定义、采集设计、数据采集与入库、标注治理、质量评测、发布共享、应用反馈与版本维护的全生命周期数据质量治理机制。与通用数据治理流程相比,农业数据集治理的关键在于将农业专业知识、质量控制节点和质量凭证嵌入数据形成与使用全过程,使数据集具备可检查、可追溯和可持续修复能力。图3展示了这一全生命周期数据质量治理与逆向溯源框架。

在任务定义阶段,应通过任务说明书明确数据集的目标用户、适用作物、适用区域、适用生育期、评价指标和限制场景,避免数据集建设目标泛化。在采集设计阶段,应围绕区域、年份、季节、生育期、品种、栽培制度、气候区和设备条件开展分层采样,形成采集方案和样本覆盖清单。在数据采集与入库阶段,应同步记录采集时间、地点、设备参数、气象条件、作物品种、生育期、管理措施等农业过程级元数据,并对格式一致性、缺失值、重复样本和异常样本进行检查。在标注治理阶段,应建立标注规范、专家复核和争议样本处理机制,记录标注批次、复核比例、专家一致率和修改过程。在质量评测阶段,除常规随机

测试外，还应设置跨区域、跨年份、跨生育期或跨设备条件下的独立测试集，以检验数据集在真实农业场景中的泛化能力。在发布共享阶段，应配套数据说明书、数据字典、授权协议、脱敏说明、DOI 和版本号，并根据数据敏感程度实施分级开放。在应用反馈与版本维护阶段，应持续收集用户反馈、失败样本和模型失效场景，并将其转化为数据补充、标注修订、元数据完善和重新评测的依据。

质量凭证是全生命周期数据质量治理机制得以落实的核心。本文所称质量凭证，是指能够记录数据集建设过程、质量状态和修订依据的结构化证据，包括任务说明书、采集方案、样本覆盖清单、采集日志、数据字典、农业过程级元数据、标注规范、专家复核记录、质量评测报告、跨域评测结果、数据说明书、用户反馈记录和版本日志等。通过为样本、批次、标注结果、元数据记录和评测结果设置统一标识，质量凭证能够在不同治理环节之间流转，为质量审核、责任追踪和版本修复提供依据。

当应用反馈显示模型在特定区域、年份、生育期或设备条件下表现下降时，可按照“应用失效/失败样本—样本 ID 追踪—元数据核查—标注批次核查—采集设计核查—缺陷类型判定—补采样/重标注/重新评测—新版本发布”的路径开展逆向溯源。若失败样本

集中于某一区域或生育期，说明采集设计可能存在覆盖不足；若失败样本集中于某一标注批次，则可能源于标注规范不清或专家复核不足；若失败样本缺少生育期、气象条件或设备参数记录，则说明农业过程级元数据不完整。由此，应用反馈能够反向驱动数据补充、标注修订和版本更新，使数据集从静态发布资源转向持续演化的质量治理对象。

以田间病虫害识别数据集为例，若早期数据主要来自实验室背景下的叶片图像，且样本集中于典型症状阶段，模型可能在内部测试中表现较好，但在复杂田间背景、不同光照条件、病害初发阶段或不同生育期中出现误诊。基于上述框架，可通过失败样本 ID 回溯其采集地点、生育期、拍摄设备、病害阶段和标注批次，从而判断问题来自样本覆盖不足、标注规范不清，还是元数据记录缺失。相应修复措施包括补充田间复杂背景样本、细化病害类型与病程阶段标注、增加专家复核，并重新开展跨区域、跨生育期和跨设备评测。该案例表明，全生命周期数据质量治理框架不仅能够描述数据集建设流程，还能够通过质量凭证实现应用失效后的逆向诊断和质量修复。

围绕上述框架，表 6 将农业数据集全生命周期数据质量治理过程进一步转化为可检查的控制节点、质量凭证和修复方式。

表 6 农业数据集全生命周期质量控制节点、质量凭证与修复方式

Table 6 Quality control checkpoints, quality evidence, and remediation approaches across the full lifecycle of agricultural datasets

生命周期阶段	质量控制节点	质量凭证	修复方式
任务定义	任务目标、适用场景、评价指标和限制场景	任务说明书、适用边界说明	修订任务范围和限制场景
采集设计	区域、年份、生育期、品种、栽培制度和设备覆盖	采集方案、样本覆盖清单	补充缺失区域、生育期或设备条件样本
数据采集与入库	格式、缺失值、异常样本和元数据完整性	采集日志、数据字典、农业过程级元数据	数据清洗、异常修正、补充元数据
标注治理	标注规范、专家一致率、复核比例和争议样本处理	标注规范、专家复核记录、争议样本记录	修订规范、重新标注、专家仲裁
质量评测	准确性、完整性、代表性和跨域泛化性	质量评测报告、跨域评测结果	调整测试划分，增加跨域测试集
发布共享	数据说明、授权协议、脱敏处理、版本号和引用方式	数据说明书、授权协议、DOI、版本号	完善说明文档，调整共享权限
应用反馈与维护	用户反馈、失败样本、失效场景和版本更新	用户反馈记录、失败样本清单、版本日志	补采样本、修订标注、重新评测、发布新版本

在开放共享与持续维护环节，不同农业数据的敏感程度、权属结构和复用条件存在差异，因而需要采

用分级开放策略。表 7 归纳了农业数据集的主要开放共享类型及其治理重点。

表 7 农业数据集开放共享类型、治理重点与代表场景

Table 7 Open-sharing types, governance priorities, and representative scenarios for agricultural datasets

开放类型	适用数据	治理重点	代表场景
完全开放	低敏科研数据	可下载、可引用、可复现	公开病虫害图像、通用气候土壤数据
脱敏开放	生产过程数据	兼顾复用价值与主体权益	农机作业记录、地块管理记录
授权访问	高敏空间或经营数据	控制用途、权限和再分发边界	高精度地块边界、农户经营数据
持续服务	动态数据产品	维护版本、质量置信和反馈机制	作物制图产品、农业资源环境数据产品

总体来看，农业高质量数据集标准框架的核心不只是建立一般化建设流程，而是将农业任务、专业标注、过程级元数据、跨域评测和分级开放机制贯穿数据全生命周期。通过质量凭证流转和逆向溯源机制，农业数据集建设能够从项目化资源积累转向证据化、可追溯、可修复的长期治理能力建设，从而更好支撑农业人工智能模型研发、数字农业平台建设、农业生产管理优化和宏观决策分析。

6 总结

农业领域高质量数据集是支撑数字农业、智慧农业和农业人工智能可靠发展的基础资源。本文系统梳理了农业高质量数据集的核心内涵与评价维度，探讨了其典型类型及面临的主要问题，并指明了未来建设路径，同时对农业数据集从资源汇聚走向全生命周期数据质量治理的必要性进行了分析。在内涵层面，高质量农业数据集并不等同于样本规模扩大，而是体现为数据真实可信且任务边界清楚，样本覆盖充分的同时确保标签准确一致；此外，元数据需完整规范并支持标准互操作，最终实现便捷的开放复用与可行的持续治理^[1-4,10,17,19,21]。

在建设现状方面，本文综述了农业视觉与病虫害、遥感与资源环境，以及文本知识与多模态生产管理等代表性数据集，并围绕场景覆盖与样本代表性、标注质量与元数据规范，以及开放共享与持续维护等多个维度进行了归纳分析。研究发现，当前农业数据集整体呈现出“数量增长较快、质量证据不足”“单点场景较多、跨域泛化不足”“模型训练导向明显、长期治理机制薄弱”等特点。尤其在应对复杂农业生产场景、开展跨时空泛化评测，以及落实专家标注复核与敏感数据分级开放等方面，仍存在较为突出的建设短板^[5,8-9,18,35,41-43,47]。

展望未来，农业高质量数据集建设应朝着“任务

清晰、质量优先、标准统一、开放分级、持续维护”的方向发展。遥感、无人机与近景图像等多源视觉信息的汇聚，结合物联网传感、农机作业与土壤气象等环境记录，再辅以农事管理及专业文本知识的深度融合，将推动农业数据集从单一训练资源转向支撑生产诊断、装备控制、区域监测与智能决策的数据基础设施。WorldCereal、AI4Boundaries 和 SoilGrids 等案例表明，引入开放源码与质量置信信息、量化空间不确定性，以及实现数据的持续更新与 AI-ready 表达，正逐渐成为农业数据产品的重要发展方向^[39,50-51]。只有构建起高质量与广覆盖兼备，且具有强互操作性及可维护性的农业数据资源体系，才能为农业科技创新、农业现代化以及粮食安全治理提供坚实支撑。

参考文献

[1] Wolfert S, Ge L, Verdouw C, et al. Big data in smart farming—a review. *Agricultural Systems*, 2017, 153: 69-80.

[2] Liakos K G, Busato P, Moshou D, et al. Machine learning in agriculture: A review. *Sensors*, 2018, 18(8): 2674.

[3] Basso B, Antle J. Digital agriculture to design sustainable agricultural systems. *Nature Sustainability*, 2020, 3(4): 254-256.

[4] 中国农业科学院农业信息研究所. 农业大数据研究进展（2024）. 农业大数据学报, 2024, 6(4): 433-468.

Agricultural Information Institute of CAAS. Progress of agricultural big data research (2024). *Journal of Agricultural Big Data*, 2024, 6(4): 433-468.

[5] Chamorro-Padial J, García R, Gil R. A systematic review of open data in agriculture. *Computers and Electronics in Agriculture*, 2024, 219: 108775.

[6] Whang S E, Roh Y, Song H, et al. Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal*, 2023, 32: 791-813. <https://doi.org/10.1007/s00778-022-00775-9>.(arXiv:2112.06409)

- [7] Paullada A, Raji I D, Bender E M, et al. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2021, 2(11): 100336.
- [8] 管博伦,张立平,朱静波,等. 农业病虫害图像数据集构建关键问题及评价方法综述. *智慧农业(中英文)*, 2023, 5(3): 17-34.
GUAN B L, ZHANG L P, ZHU J B, et al. The key issues and evaluation methods for constructing agricultural pest and disease image datasets: A review. *Smart Agriculture*, 2023, 5(3): 17-34.
- [9] 姚艳敏,白玉琪. 农业大数据标准体系框架研究. *农业大数据学报*, 2019, 1(4): 76-85.
Yao Y M, Bai Y Q. A Framework for agricultural big data standards. *Journal of Agricultural Big Data*, 2019, 1(4): 76-85.
- [10] Wang R Y, Strong D M. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 1996, 12(4): 5-33.
- [11] Batini C, Cappiello C, Francalanci C, et al. Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 2009, 41(3): 1-52.
- [12] Cai L, Zhu Y Y. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 2015, 14: 2.
- [13] Strong D M, Lee Y W, Wang R Y. Data quality in context. *Communications of the ACM*, 1997, 40(5): 103-110.
- [14] Wang R Y. A product perspective on total data quality management. *Communications of the ACM*, 1998, 41(2): 58-65.
- [15] Lee Y W, Strong D M, Kahn B K, et al. AIMQ: A methodology for information quality assessment. *Information & Management*, 2002, 40(2): 133-146.
- [16] Saaty T L. *The Analytic Hierarchy Process*[M]. New York: McGraw-Hill, 1980.
- [17] Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 2016, 3(1): 1-9.
- [18] Ali B, Dahlhaus P. The role of FAIR data towards sustainable agricultural performance: a systematic literature review. *Agriculture*, 2022, 12(2): 309.
- [19] Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *Communications of the ACM*, 2021, 64(12): 86-92.
- [20] Pushkarna M, Zaldivar A, Kjartansson O. Data cards: Purposeful and transparent dataset documentation for responsible Ai[C]//*Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022: 1776-1826.
- [21] Yue P, Shangguan B, Ziebelin D. The OGC training data markup language for artificial intelligence (TrainingDML-AI) standard[C]//*EGU General Assembly Conference Abstracts*. 2023: EGU-16998. DOI:10.5194/egusphere-egu23-16998.
- [22] Cheng C, Messerschmidt L, Bravo I, et al. A general primer for data harmonization. *Scientific Data*, 2024, 11(1): 152.
- [23] Didwania K, Seth P, Kasliwal A, et al. AgriLLM: Harnessing transformers for farmer queries[C]//*Proceedings of the Third Workshop on NLP for Positive Impact*. Miami, Florida, USA. Stroudsburg, PA, USA: ACL, 2024: 179-187.
- [24] Yang S, Liu Z, Mayer W, et al. ShizishanGPT: An agricultural large language model integrating tools and resources[C]//*Web Information Systems Engineering - WISE 2024. Lecture Notes in Computer Science*, vol 15439. Springer, Singapore, 2025: 284-298. https://doi.org/10.1007/978-981-96-0573-6_21. (arXiv:2409.13537)
- [25] Awais M, Alharthi A H S A, Kumar A, et al. AgroGPT: Efficient agricultural vision-language model with expert tuning[C]//*2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Tucson, AZ, USA, 2025:5687-5696. DOI: 10.1109/WACV61041.2025.00555. (arXiv:2410.08405)
- [26] Hu Z Q, Wang L, Lan Y H, et al. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models[C]// *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore. Stroudsburg, PA, USA: ACL, 2023: 5254-5276.
- [27] Gao Y, Xiong Y, Gao X, et al. Retrieval-Augmented Generation for Large Language Models: A Survey[PP/OL]. 2023, arXiv:2312.10997.
- [28] Hughes D, Salathé M. An open access repository of images on plant health to enable the development of mobile disease diagnostics [PP/OL]. 2015, arXiv:1511.08060.
- [29] Mohanty S P, Hughes D P, Salathé M. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 2016, 7: 1419.
- [30] Singh D, Jain N, Jain P, et al. PlantDoc: A dataset for visual plant disease detection[C]//*Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. Hyderabad India. ACM, 2020: 249-253.
- [31] Wu X P, Zhan C, Lai Y K, et al. IP102: A large-scale benchmark dataset for insect pest recognition[C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*. 2019: 8779-8788. DOI: 10.1109/CVPR.2019.00899.
- [32] Olsen A, Kononov D A, Philippa B, et al. DeepWeeds: A multiclass weed species image dataset for deep learning. *Scientific Reports*, 2019, 9(1): 2058.
- [33] David E, Madec S, Sadeghi-Tehran P, et al. Global wheat head

- detection (GWHD) dataset: A large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, 2020: 3521852.
- [34] Chiu M T, Xu X Q, Wei Y C, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2825-2835.
- [35] 王伯元,管志浩,杨杨,等. 2023 年南京农业病虫害图像识别数据集. *农业大数据学报*, 2023, 5(2): 91-96.
- Wang B Y, Guan Z H, Yang Y, et al. The agricultural pest and disease image recognition dataset in Nanjing, Jiangsu Province, in 2023. *Journal of Agricultural Big Data*, 2023, 5(2): 91-96.
- [36] Tseng G, Zvonkov I, Nakalembe C, et al. CropHarvest: A global dataset for crop-type classification[C]//NeurIPS Datasets and Benchmarks. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). 2021. <https://api.semanticscholar.org/CorpusID:248529758>.
- [37] Sykas D, Sdraka M, Zografakis D, et al. A sentinel-2 multiyear, multicountry benchmark dataset for crop classification and segmentation with deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022, 15: 3323-3339.
- [38] Schneider M, Schelte T, Schmitz F, et al. EuroCrops: The largest harmonized open crop dataset across the European Union. *Scientific Data*, 2023, 10(1): 612.
- [39] Poggio L, De Sousa L M, Batjes N H, et al. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil*, 2021, 7(1): 217-240.
- [40] Fick S E, Hijmans R J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 2017, 37(12): 4302-4315.
- [41] Wu H, Xie N F, Wang X L, et al. Crop GraphRAG: pest and disease knowledge base Q&A system for sustainable crop protection. *Frontiers in Plant Science*, 2026, 16: 1696872.
- [42] Yao X C, Hao X, Liu R L, et al. AgCNER, the first large-scale Chinese named entity recognition dataset for agricultural diseases and pests. *Scientific Data*, 2024, 11(1): 769.
- [43] 杜林煊,曹姗姗,刘婷婷,等. 奶牛个体头部多模态图像深度学习训练数据集. *中国科学数据(中英文网络版)*, 2025, 10(1): 11-20.
- Du L X, Cao S S, LIU T T, et al. A dataset of deep learning training for multimodal headimagesofindividual cows. *China Scientific Data*, 2025, 10(1): 11-20.
- [44] 池美香,陈韶萍,黄婷,等. 柑橘黄龙病田间症状识别图像数据集. *中国科学数据(中英文网络版)*, 2025, 10(1): 52-61.
- Chi M X, Chen S P, Huang T, et al. A dataset of field symptoms for identifying Citrus Huanglongbing. *China Scientific Data*, 2025, 10(1): 52-61.
- [45] 张雨,饶元,陈文骏,等. 不同成熟度番茄果实多模态图像数据集. *中国科学数据(中英文网络版)*, 2025, 10(1): 73-88.
- Zhang Y, Rao Y, Chen W J, et al. A dataset of multimodal images of tomato fruits at different stagesof maturity. *China Scientific Data*, 2025, 10 (1): 73-88.
- [46] Northcutt C G, Athalye A, Mueller J. Pervasive label errors in test sets destabilize machine learning benchmarks[C]//Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021) round 1. (arXiv:2103.14749)
- [47] Kindling M, Strecker D. Data quality assurance at research data repositories. *Data Science Journal*, 2022, 21: 18.
- [48] Van Tricht K, Degerickx J, Gilliams S, et al. WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping. *Earth System Science Data*, 2023, 15(12): 5491-5515.
- [49] Tonmoy S M T I, Zaman S M M, Jain V, et al. A Comprehensive Survey of hallucination mitigation techniques in large language models[PP/OL]. 2024, arXiv:2401.01313.
- [50] d'Andrimont R, Claverie M, Kempeneers P, et al. AI4Boundaries: an open AI-ready dataset to map field boundaries with Sentinel-2 and aerial photography. *Earth System Science Data*, 2023, 15(1): 317-329.
- [51] Hengl T, Mendes de Jesus J, Heuvelink G B M, et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, 2017, 12(2): e0169748.

引用本文: 谢能付,王晓丽,王馨,孟智博,满芮,赵忠贺,吴昊,吕前杰. 农业高质量数据集评价框架与全生命周期治理路径[J]. *农业大数据学报*, 2026, 8(3): 354-372. DOI: 10.19788/j.issn.2096-6369.000168.

CITE: XIE NengFu, WANG XiaoLi, WANG Xin, MENG ZhiBo1, MAN Rui, ZHAO ZhongHe, WU Hao, LV QianJie. Agricultural High-Quality Dataset Evaluation Framework and Full- Lifecycle Governance Pathway[J]. *Journal of Agricultural Big Data*, 2026, 8(3): 354-372. DOI: 10.19788/j.issn.2096-6369.000168.

Agricultural High-Quality Dataset Evaluation Framework and Full-Lifecycle Governance Pathway

XIE NengFu^{1,2,3}, WANG XiaoLi^{1,2,3*}, WANG Xin^{1,2,3*}, MENG ZhiBo^{1,2,3}, MAN Rui^{1,2,3}, ZHAO ZhongHe^{1,2,3}, WU Hao^{1,2,3}, LV QianJie^{1,2,3}

1.Agricultural Information Institute of CAAS, Beijing 100081, China; 2.National Agriculture Science Data Center, Beijing 100081, China; 3.DataHub, Chinese Long-term Experimental Network for Agricultural and Rural Development, Beijing 100081, China

Abstract: High-quality agricultural datasets are fundamental resources that support the development of intelligent agricultural models, digital agriculture applications, and scientific decision-making. The effectiveness of their construction depends heavily on standardized data collection, professional annotation, and full-lifecycle data quality governance. Most existing agricultural datasets are built for specific model training and localized scenarios. They lack a systematic framework designed for complex agricultural environments, cross-source data interoperability, and sustained open reuse. This paper systematically reviews the core concepts and evaluation dimensions of high-quality agricultural datasets. It provides a focused overview of the current construction status of datasets related to agricultural computer vision and pest control, agricultural remote sensing and environmental resources, as well as agricultural textual knowledge and multimodal production management. The study first explains the foundational role and quality evaluation logic of high-quality datasets in subsequent agricultural artificial intelligence development and digital agriculture applications. It then analyzes the characteristics and limitations of existing agricultural datasets regarding requirement boundaries, sample representativeness, annotation reliability, metadata specifications, open sharing, and compliance governance. Finally, the paper outlines future construction directions from three dimensions encompassing task-driven design, full-lifecycle data quality governance, and sustainable open reuse. Future efforts need to prioritize real-world scenario coverage, data quality governance, the implementation of FAIR principles, and AI-ready representations. The study also suggests exploring optimization mechanisms for version management, application feedback, and multi-stakeholder collaboration. These insights aim to provide theoretical references and practical guidance for building a reliable and robust data supply system that meets the specific demands of agricultural scenarios.

Keywords: agricultural datasets; data quality; Ai-ready; quality evaluation; full-lifecycle data quality governance; agricultural artificial intelligence; open sharing