

涉农数据质量维度研究

靳守业 卢月静 张晓玲 高思宇 贾启山

北京市数字农业农村促进中心

DOI:10.12238/as.v8i5.3022

[摘要] 本文主要阐述了涉农数据质量在智慧农业、数字乡村建设等“三农”工作中的重要作用,分析了涉农数据在采集、传输、存储、应用全生命周期存在的质量问题,参照相关标准对相关数据质量维度进行分析与测量,旨在通过提升涉农数据质量促进数据高效应用,进而有效提升三农工作质量和效率。

[关键词] 涉农数据; 质量维度; 智慧农业

中图分类号: DF413.1 **文献标识码:** A

Study on the quality dimension of agriculture-related data

Shouye Jin Yuejing Lu Xiaoling Zhang Siyu Gao Qishan Jia

Beijing Digital Agriculture and Rural Development Promotion Center

[Abstract] This paper primarily elucidates the critical role of agricultural data quality in smart agriculture and digital rural construction, among other "three-rural" initiatives. It analyzes the quality issues that exist throughout the entire lifecycle of agricultural data, from collection to transmission, storage, and application. By referring to relevant standards, it examines and measures various dimensions of data quality. The aim is to enhance the quality of agricultural data to promote its efficient use, thereby effectively improving the quality and efficiency of "three-rural" work.

[Key words] agriculture-related data; quality dimension; smart agriculture

引言

数据主要用来描述事物特征、特征和过程变化等现象,涉农数据描述内容包括农业生产状况(农作物种植面积、牲畜存栏量和肉奶产量等),资源利用(土地资源类型、面积、肥力),农村经济(农户收入结构、农村产业结构),农村基础民生(农村人口数量、教育水平)等。涉农数据已成为驱动“三农”工作数字化转型的关键要素。从完整性、一致性、时效性等重要质量维度对涉农数据进行分析测量和问题整改,对提升涉农数据质量和“三农”工作水平具有十分重要的作用。

1 涉农数据质量常见问题

涉农数据质量问题贯穿于数据全生命周期的各个阶段,主要包括标准不统一、安全隐患、利用深度不足等。

1.1 数据采集环节常见

涉农数据采集主体众多,不同采集主体制定的数据采集标准规范在统计口径、采集方法等方面存在差异。例如,在估算蔬菜或粮食产量时,有的部门按照单位面积农作物产量来估计总产量,有的部门按照农作物单株产量来估计总产量,造成数据可靠性不足。偏远地区因网络覆盖不足等原因易造成数据时效性差。此外,农户文化素质和信息化水平参差不齐,也易造成其采集数据的质量高低不一。

1.2 数据存储环节常见问题

政府单位、集体组织和科研机构建设和使用的涉农信息系统数据相互隔离,难以实现共享和整合。同时农户个人信息、农业科技信息等涉农敏感信息面临数据泄露和篡改等安全风险。

1.3 数据应用环节常见问题

涉农数据的应用效果和深度受数据质量显著影响。例如自动驾驶农机依赖高精度农田数据支撑,若误差大会直接导致作业效率降低。

2 涉农数据质量评测分析

为充分发挥数据价值,高效利用涉农数据,需对涉农数据的质量维度进行准确分析、测量和问题整改来有效提升数据质量。

2.1 数据质量维度分析

许多学者深入研究了高质量数据评估维度、方法和架构等。Strong-Wangs架构提出数据质量应重点从固有质量维度、场景、表达、访问数据四个维度进行提升,Tommas Redman提出基于数据结构的维度进行数据质量分析,该维度主要与数据的模型(实体和属性)的值组成部分相关联,共描述了关联性、同一性、最小必要冗余性等24个维度。《信息技术 数据质量评价指标》(GB/T36344-2018)基于安全角度将规范性、准确性、完整性、一致性、时效性和合规性作为数据质量评价指标。数据管理协会

表 1 涉农数据质量维度

数据生成 质量维度	数据采集 质量维度	数据传输 质量维度	数据存储 质量维度	数据处理 质量维度	数据共享 质量维度	数据应用 质量维度
1.准确性 数据反映“三农”工作实 情,包括农产品产量等准 确信息。	1. 代表性 采集的数据能够代 表“三农”工作的全 部情况。	1. 安全性 采取有效的安全措 施保障数据安全传 输。	1. 安全性 采取有效的安全措 施防止发生数据泄 漏、损坏等问题。	1. 准确性 确保在清洗、转换等处 理过程中未发生数据 失真问题。	1. 安全性 确保数据共享过程不 发生篡改、丢失等安 全问题。	1. 有效性 应用后能发挥提高生产 效率、为政策制定提供有 效支持的效果。
2.完整性 生产量、生产地、农作物 品种等涉农数据应记录 完整。	2. 时效性 及时采集数据,减少 数据滞后影响。	2. 完整性 传输过程未丢失或 损坏数据。	2. 可靠性 存储环境安全可 靠,确保数据安全 稳定存储。	2. 完整性 保证数据完整,不丢失 或损坏数据。	2. 完整性 保持内容齐全、结构 完整、逻辑一致。	2. 易用性 应用后便于数据使用者 使用和分析,便于产生价 值。
3.一致性 数据生成过程中应基于 相同的统计口径、编码和 定义规则等标准。		3. 高效性 重点关注传输速度 和资源使用效率,包 括带宽利用率、丢包 率等。	3. 可扩展性 数据量在快速增长 过程中,存储系统 能及时扩展,确保 正常使用。	3. 合规性 数据处理要符合法律 法规要求,同时要做好 权限划分,数据脱敏。	3. 可靠性 共享措施应保持稳 定,确保数据能持续、 稳定共享。	3. 价值性 通过对涉农数据深度挖 掘,实现利用最大化。
		4. 时效性 数据及时传输到目 的地。	4. 完整性 数据未丢失、损坏 或遭篡改。			
1. 可解释性。明确数据来源的方式、标注采集数据方法和范围,详细描述数据清洗过程和方法、模型等,清晰明确变量和指标,采用合适的指标和方法评估数据质量。						
2. 可追溯性。对涉农数据生命周期,对数据来源、采集过程、处理步骤、存储位置和使用情况等信息进行完整记录,便于及时追溯到数据的任何一个环节。						
3. 关联性。涉农各领域数据存在的相互联系和依存关系。						
4. 可测量。制定方案和标准对涉农数据质量进行量化评估。以便准备掌握数据质量问题。						

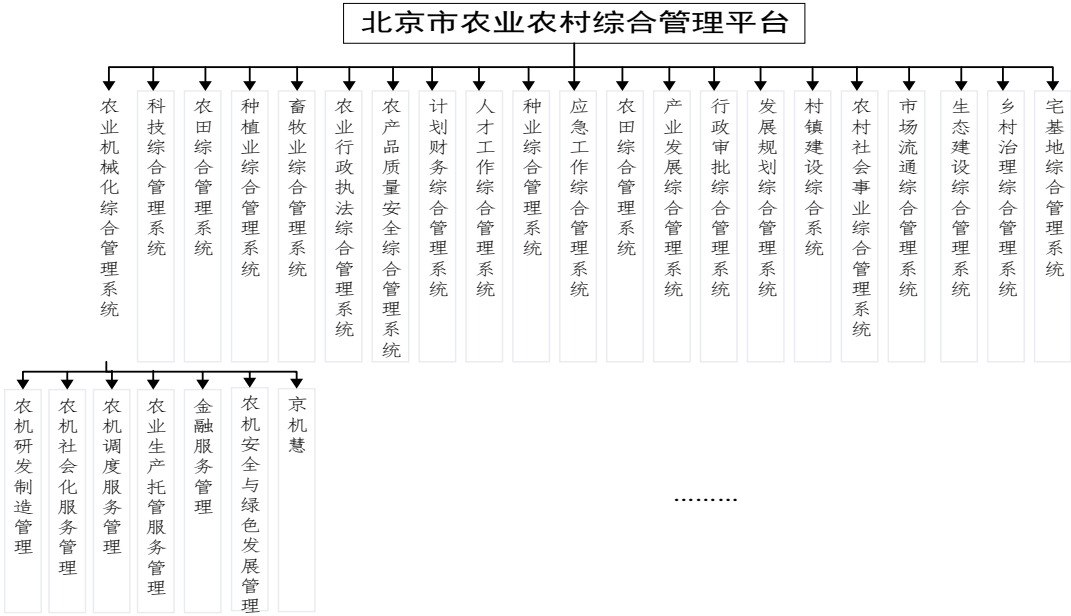


图 1 农业农村综合管理平台架构

(DAMA)描述数据质量包括6个核心维度和5个普通特征,其中核心维度包括完备性、唯一性、及时性、有效性等,普通特征包括可用性、时间问题、灵活性和置性度。以上分析方法存在的问题包括: Strong-Wangs架构评估涉农数据质量存在忽视安全性需求的问题,基于安全维度评估涉农数据质量存在安全标准不通用,对场景应用效果和表达等方面重视不够的问题。基于数据结构的质量评估方法存在评估维度单一、维度之间互补性不足等问题。DAMA定义数据质量维度存在定义模糊、逻辑关系复杂和缺乏涉农业务针对性等问题。

2. 2涉农数据质量维度分析

为使涉农数据更适合“三农”工作的开展,需综合数据结构、业务流程等因素,对数据由产生到销毁的全生命周期对数据质量维度进行分析。重点维度包括准确性、完整性、一致性、可追溯性、可解释性、相关性等。

表 2——数据测量内容和方法

序号	属性	测量方法	测量内容
1	完整性	1. 哈希值校验 2. 冗余值校验 3. 取值一致性检查	选取完备数据集,对数据进行哈希值校验来确定是否被篡改。通过对发送和接收到的数据进行冗余值校验来研判传输和存储数据是否错误。检查数据在不同存储位置、不同业务系统的取值,确保同一数据取值一致。
2	准确性	1.采集阶段,采集数据与入库数据对比。2.清洗阶段。规则校验,逻辑一致。3.业务验证。管理员手动验证和修改错误数据。	1.逻辑一致。采集不同系统存储的同一数据,交叉验证。2.抽样验证。从数据集中随机选取有代表性的数据样本,对样本进行准确性验证。3.外部数据验证。依据第三方统计数据验证。4.脏数据检测。通过对数据逻辑检测等方式确认是否有空数据、过期数据、格式不一致的数据等。
3	一致性	1.不同数据源采集的数据的格式、编码、类型保持一致。2.存储和使用时参照准确性和完整性标准。3.数据在清洗、转换等处理过程中的逻辑正确、结果一致。4.销毁数据时要验证与采集数据一致。	1.完整性检查。验证数据表中的关键字段是否具有唯一性。检查外键值在关联表中是否存在对应的主键值。确认必填字段是否都有值,确保不存在空值问题。2.逻辑检查。检查同一数据集内不同数据元素之间的逻辑关系是否一致。检查数据与其他相关系统或数据源中的对应数据保持一致。3.数据一致性和时间戳检查。验证数据是否符合格式要求,检查有时间参数的数据的时间戳顺序是否一致。
4	时效性	1.基于时间戳的方法。计算当前时间与时间戳的差值。2.统计更新次数,在一定时间段内,统计数据的更新次数。3.业务规则法,根据业务需求确定数据在不同时间的价值。	1.通过程序确定数据的采集时间、更新时间 and 有效时间,确认数据是否及时得到了修正和补充。2.检测数据完整性,确认是否有必填项出现空值。3.数据一致性,检测多源数据的时间是否一致,长期数据是否可回溯且无矛盾,是否有过时数据。
5	可追溯性	1.数据全生命周期均有可追溯记录。2.数据全生命周期数据完整性、准确性、一致性符合质量要求。	1.确认数据采集源,重点通过传感器、问卷、数据库记录等记录来源信息。2.数据整合和关联,记录整合方式和关联规则、标识等。3.传输过程记录,记录传输路径、传输协议、加密算法、传输时间等记录。4.使用共享记录。记录共享对象、共享时间、共享范围、共享方式等记录
6	可靠性	1.检测数据完整性、准确性、一致性。2.数据稳定性。选取一段时间,检查数据波动是否在合理范围。3.数据异常性检测。检测数据中是否有异常值。	1.检测数据是否严格审核和验证。 2.数据波动是否在合理范围,异常值在可控范围内并不会造成较大影响。
7	可解释性	1.可追溯性验证。2.数据逻辑性验证,内部逻辑一致,与现实逻辑相符。3.数据可理解性验证。数据表示易懂性、含义明确,含义清晰明确。	1.复杂度评估,通常维度越高、特征数量越多且相关性越强,复杂度越高。2.语义清晰度评估。字段和变量定义明确、清晰,数据在不同部分、不同数据集中语义保持一致。3.问卷调查。对用户进行问卷调查,看用户对数据理解是否存在困难和误解。
8	可扩展性	1.技术架构可扩展性测量。 2.业务场景可扩展性测量。	1.评估存储系统能否通过增加节点、磁盘、云存储资源等或所采用的分布式架构,实现扩展存储容量并且不影响对数据和各种操作。2.数据模型是否支持动态字段扩展。
9	合规性	操作符合《数据安全法》《个人信息保护法》等法律法规。	1.数据分类分级 2.遵守“最小授权原则”和“权责分离原则”
10	开放性	在可获取性、可访问性、合规性等方面的开放程度。	参照可访问性、可理解性、合规性的测量内容。
11	易用性	数据可访问性、可理解性、可操作性、兼容性、时效性、文档与支持等。	1.用户获取数据的便捷程度,包括访问渠道、权限透明度和获取成本。(权限设计完善、API 数据交互是否高效等)。2.数据语义清晰、可视化程度。3.数据是否便于用户分析、处理和二次开发。 4.相关的辅助资源和技术支持是否有效降低用户学习成本。
12	价值性	数据对“三农”的应用价值、决策支持能力及经济效益。	1.对农业生产效率提升效果。2.对决策制定的支持效果。 3.经济价值、社会效益、用户满意度等。
13	安全性	测试数据全生命周期安全性能。	评估工作围绕数据资产、安全威胁、脆弱性、风险 4 个方面开展测评,识别可能对数据资产造成损害的潜在原因。

2. 3 涉农数据质量维度设计

参照数据生产者、使用者等对数据质量需求和“三农”工作业务特点,本文面向涉农数据全生命周期规划了数据在不同阶段的质量维度,如表1所示:

2. 4 涉农数据质量测量方法和测量内容

数据质量测量工作主要包括数据质量需求分析、选取质量维度、确定测量维度和评价方法、运用选定方法进行质量评估、测量结果分析等步骤。

2. 4. 1 数据质量测量方法

数据质量测量方法主要是综合《信息安全技术 数据安全能力成熟度模型》《信息安全技术 个人信息安全影响评估指南》《信息安全技术 网络安全等级保护基本要求》等标准规范对涉农数据从个人信息收集的“最小必要”原则、用户获取明示同意的方式、数据加密与去标识化技术措施,数据共享、公开披露的合

规流程进行综合测量。基于《质量管理体系要求》《信息技术数据质量评价指标》等标准规范对数据开放性、可解释性、可追溯性等维度进行测量。

2. 4. 2 涉农数据质量测量内容

针对涉农数据质量开展逐项测量,对不符合标准和规范的数据质量维度进行问题整改,可有效提升数据质量,测量方法和内容如表2所示。

3 涉农数据质量提升案例

北京市农业农村综合管理平台是基于北京市农业农村局工作实际构建的综合业务管理系统。该平台整合了各类涉农数据资源,目标是通过有效的数据治理推动智慧城市建设与农业农村发展的深度融合。目前已完成21个子系统建设,如图1所示。

3. 1 数据质量测量维度确定

表 3 数据安全维度测量

主机脆弱性	数据库安全	安全漏洞	中间件安全问题
1. 文件与目录未进行权限控制。 2. 未设置日志文件大小策略。 3. 未配置用户最小授权。 3. 未设置登录超时时间。	1. 未采用两种以上组合的鉴别技术对用户进行身份鉴别。 2. 未对重要主体和客体设置安全标记。 3. 未基于可信根对计算设备的系统程序、重要配置参数和应用程序等进行可信验证。 4. 未采用校验技术或密码技术保证审计数据、配置数据存储过程中的完整性。	1. 任意密码重置 2. 未授权访问 3. xxl-job 命令执行 4. Shiro 反序列化 5. 穿越网络边界问题 6. SQL 注入 7. 未授权访问 8. 信息泄露	Tomcat 安全漏洞易造成数据泄露。

表 4—完整性等维度测量结果

完整性	准确性	一致性	时效性
1. 数据字段完整性 非空字段占比 67% 2. 数据记录完整性 因部分子系统尚未正式使用, 因此部分源数据未完全导入该系统数据库, 目前记录完整性为 83%。	1. 数据内容正确性和、格式合规性较高, 但存在大量重复数据和无用脏数据, 占比约 30%。 30%。	1. 内部数据一致, 例如粮食平均亩产量*亩数=粮食总产量。 2. 部分多源数据因采集、汇聚不完整存在部分不一致数据, 占比约 10%。	本年度内有更新的数据占比为 44%, 近三个月内更新的数据占比仅为 26%。

表 5 可追溯性等维度测量结果

可追溯性	可靠性	可解释性	可扩展性
数据从采集源头到最终存储和使用的过程中记录完整, 包括数据采集时间、采集人员、采集设备等信息。数据在传输、存储和处理过程中记录每一步操作和转换的日志, 日志存储时间超过 6 个月。	参照完整性和一致性、数据来源稳定性来评估可靠性。经检查, 不同来源或不同时间采集的同一涉农数据存在部分不一致问题, 例如不同批发市场上报农产品市场行情价格分布范围过大。	1. 建立元数据管理功能增强数据可解释性, 但是元数据设计不完善, 可解释性不足。 2. 70% 的数据库字段未添加描述。	1. 系统管理基于支持应用平台, 新建子系统的定义和文档较完整性, 新增数据来源与原数据配备较复杂, 可扩展能力不高。 2. 系统性能和容量可扩展性良好, 新增磁盘空间效率较高。预留了足够的磁盘空间和计算资源等。

表 6 合规性等维度测量结果

合规性	开放性	易用性	价值性
1. 数据处理过程符合《网络安全法》《数据安全法》等法律法规要求。 2. 收集个人信息进行明示, 符合《个人信息保护法》要求。 3. 数据传输、存储均操作经国密加密算法 SM4 加密的数据。	1. 数据开放审批手续较完善。 2. 由于可解释性较弱, 开放可使用性较弱。 3. 农业经济数据等部分数据更新不及时影响开放效果。	1. 数据完整性、数据相关性不足对可用性造成影响。 2. 界面设计有助于数据的易用性, 但更新不及时影响易用性效果。	1. 农业机械、农经综合等 4 个领域, 数据挖掘与利用不断加强, 领域驾驶舱设计和建设已初步完成, 涵盖 9 个专题和 51 类指标。 2. 数据治理效果不明显造成应用价值不高。

基于本文提出的质量维度应从安全性、完整性、准确性、一致性、可追溯性、可靠性、可解释性、可扩展性、合规性、开放性、易用性、有价值等维度进行逐一测量。

3. 2 农业农村综合管理平台质量测量

3. 2. 1 安全风险评估

通过漏洞扫描、代码审计、主机安全检测等措施发现数据安全方面的问题, 其中数据库自身存在的问题包括未采用双因子验证技术对用户进行身份鉴别等, 详情见表 3 所示。

3. 2. 2 完整性、准确性、一致性等重要维度质量测量结果
利用质量维度测量工具和方法, 得到结果主要包括非空数

据占比较高、存在脏数据, 详细情况如表 4 所示:

3. 2. 3 可追溯性、可靠性、可解释性、可扩展性等维度质量测量结果

可追溯性等维度质量测量结果主要包括可解释性和可靠性不足, 可追溯性和可扩展性较好, 如表 5 所示。

3. 2. 4 合规性、开放性、易用性、价值性等质量维度测量
对合规性等质量维度测量得到结果包括处理过程合规, 但受完整性不足等原因影响易用性、开放效果和数据应用价值不高, 如表 6 所示。

3. 3 数据质量提升策略

农业农村综合管理平台数据准确性、时效性、一致性、合规性等维度质量较高,完整性、易用性和价值性等维度质量有所欠缺,为提升涉农数据价值和品质,应从完善管理制度、统一数据标准等方面开展。

3.3.1完善管理制度

建立并完善管理制度。对数据操作作出明确制度规定。为涉农数据建立质量档案,记录数据操作信息。确保发现数据质量问题,能够及时确认责任人。

3.3.2统一数据全生命周期操作标准

制定涉农数据操作的范围、内容、方式等标准。同时建立严格的数据审核流程和标准,对收集到的涉农数据进行多级审核校验。

3.3.3强化数据安全保障

采用数据加密技术、访问控制技术等安全技术对涉农敏感数据进行有效的安全防护,定期对涉农数据进行全量和增量备份,切实保障数据的可用性。

3.3.4强化涉农数据融合应用

强化数据在“三农”工作的融合应用,期间建立并逐步完善质量监控指标体系和数据品质提升的长效机制,确保通过数据融合应用促进数据品质的持续提升。

4 结论

通过涉农数据的完整性、准确性、一致性、时效性、可追溯性、可靠性、可解释性、可扩展性、合规性、安全性等质量维度进行分析和测量,在评估的基础上进行问题整改可有效提升涉农数据品质,助力“三农”工作的高效开展。

[参考文献]

[1]陈明猷,罗陆锋.采摘机器人全果园视觉感知及自主作业综述[J].智慧农业(中英文),2024,6(5):20-39.

[2]孙鲁青,李婧,郭公浦.5G背景下物联网技术助推智慧农业建设模式研究——以中小林果种植企业为例[J].智慧农业导刊,2021(5):4.3

[3]董清潭.大数据时代个人信息安全保护对策[J].电脑知识与技术,2015,(29):18-20.

[4]刘红熠,杨妮妮.互联网征信背景下个人信息主体权利保护问题研究[J].征信,2016(6).

[5]杜银霞,李红睿.大数据安全与敏感数据保护技术应用实践[J].科技风,2020(14).

作者简介:

靳守业(1982--),男,汉族,山西大同人,硕士,工程师,研究方向:农业信息化。