

公共数据管理规范

第 6 部分：公共数据质量评估

Specification for the management of public data

Part 6: Public Data Quality Assessment

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX - XX - XX 发布

XXXX - XX - XX 实施

江苏省市场监督管理局 发 布

目 次

前 言	II
引 言	1
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 评估原则	2
4.1 客观性原则	2
4.2 全面性原则	2
4.3 可比性原则	2
4.4 实用性原则	2
4.5 持续性原则	2
5 评估指标体系	2
5.1 指标体系总体框架	2
5.2 评估指标及说明	3
6 评分计算方法	3
6.1 数据集质量评分模型	3
6.2 部门数据质量评分模型	4
6.3 数据集权重配置	4
6.4 评估等级划分	4
7 数据质量评估程序	4
7.1 评估流程	4
7.2 评估准备阶段	5
7.3 指标选择阶段	5
7.4 规则确定阶段	5
7.5 评估实施阶段	5
7.6 核实整改阶段	5
7.7 数据应用阶段	5
7.8 评估总结阶段	5
8 数据质量持续改进	5
附录 A	7
参 考 文 献	15

前 言

本文件按照GB/T 1.1—2020《标准化工作导则 第3部分：标准化文件的结构和起草规则》的规定起草。

本文件是DB32/T XXXX《公共数据管理规范》的第6部分。DB32/T XXXX已经发布了以下部分

- 第1部分：数据分类分级
- 第2部分：数据共享交换
- 第3部分：数据开放
- 第4部分：数据治理
- 第5部分：数据异议处理
- 第6部分：公共数据质量评估

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由江苏省数据局提出并组织实施。

本文件由江苏省数据标准化技术委员会归口。

本文件起草单位：江苏省大数据管理中心、盐城市大数据管理中心、江苏省人力资源和社会保障信息中心。

本文件主要起草人：杨波、李萍、吴善鹏、刘超、刘瑛、庄光光、花凌锋、杨坤、朱小燕、戴祥、袁瑞红、付港、薛步高、魏从静、吴炳璋、魏瑞、仝玉超、刘伟。

引 言

公共数据是公共管理和服务机构为履行法定职责、提供公共服务收集、产生的，以电子或者其他方式记录、保存的具有公共使用价值的信息记录。规范公共数据管理、保障公共数据安全，对推动数字政府建设、提升政府治理能力和公共服务水平具有重要意义。DB32/T 4608《公共数据管理规范》旨在确立公共数据管理活动准则，由以下部分组成。

- 第1部分：数据分类分级。目的在于确立公共数据分类分级的规则、方法、流程等，为数据应用和保护寻找平衡点提供指导。
- 第2部分：数据共享交换。目的在于指导开展省、市公共数据共享交换平台建设、管理和对接，规范各级公共管理和服务机构的数据接入、归集、交换共享等工作，建立健全全省一体化公共数据共享交换体系，促进全省数据高效共享和安全交换。
- 第3部分：数据开放。目的在于规范公共数据开放的原则、过程、安全等管理要求，为全省公共数据开放工作提供参考指导。
- 第4部分：数据治理。目的在于明确公共数据治理的目标、原则、架构体系、组织管理、方法流程、运行管理、风险防控等内容，为全省公共数据治理提供参考指导。
- 第5部分：数据异议处理。目的在于指导建立完善公共数据异议的“提出、受理、核实、处理、反馈、评价”的全流程闭环处理机制，实现异议数据的高效处理，在数据使用过程中不断提升数据质量和服务水平。
- 第6部分：公共数据质量评估。目的在于明确公共数据质量评估的评估原则、总体框架、指标体系、评估流程等内容与要求，指导数据质量管理相关工作，促进数据质量持续提升。

公共数据质量评估规范

第6部分：公共数据质量评估

1 范围

本文件给出了公共数据质量评估的评估原则、总体框架、指标体系、评估流程等内容与要求。

本文件适用于公共管理和服务机构、数据主管部门、公共数据运行管理机构、从事公共数据治理的相关单位或个人开展数据工作（数据治理、“高效办成一件事”、高质量数据集、人工智能数据集训练等），对数据质量评估使用和参考。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 1.1-2020 标准化工作导则 第1部分：标准化文件的结构和起草规则
GB/T 35295-2017 信息技术 大数据 术语
GB/T 34950-2017 非结构化数据管理系统参考模型
GB/T 36344-2018 信息技术 数据质量评价指标
GB/T 37988-2019 信息安全技术 数据安全能力成熟度模型
GB/T 22240-2020 信息安全技术 信息系统安全等级保护定级指南
GB/T 42138-2022 信息技术 数据质量评价过程
GB/T 42140-2022 信息技术 数据质量评价方法
GB/T 26816-2025 信息资源核心元数据
DB32/T 4608.5-2026 公共数据管理规范 第5部分：数据异议处理

3 术语和定义

下列术语和定义适用于本文件。

3.1

结构化数据 structured data

一种数据表示形式，按此种形式，由数据元素汇集而成的每个记录的结构都是一致的并且可以使用关系模型予以有效描述。

[来源：GB/T 35295-2017, 2.2.13]

3.2

非结构化数据 unstructured data

没有固定逻辑结构的数据，如文本、图像、音频、视频等。

[来源：GB/T 34950-2017, 2.1.25]

3.3

元数据 metadata

定义、描述、管理其他信息资源的数据，用于资源发现、定位、共享、质量核验。

[来源：GB/T 26816-2025, 3.2]。

3.4

数据质量 data quality

基于对数据使用的预期，在指定条件下使用时，数据的特性满足明确的和隐含的要求的程度。
[来源：GB/T 36344, 2.3]。

4 评估原则

4.1 客观性原则

评估应基于客观事实和可验证的数据，避免主观臆断。

4.2 全面性原则

评估应覆盖数据质量的主要维度和关键指标，确保评估结果的完整性。

4.3 可比性原则

评估标准和方法应具有统一性和规范性，确保不同数据集之间的评估结果具有可比性。

4.4 实用性原则

评估指标和方法应具有可操作性，能够指导实际的数据质量改进工作。

4.5 持续性原则

数据质量评估应作为持续改进的过程，定期开展并跟踪改进效果。

5 评估指标体系

5.1 指标体系总体框架

本数据质量评估框架如图 1 所示，聚焦数据质量基础维度，包含数据完整性、准确性、一致性、及时性、有效性、唯一性六项一级指标；各项一级指标分别配置对应二级量化指标，构建分层分级的数据质量评估体系。

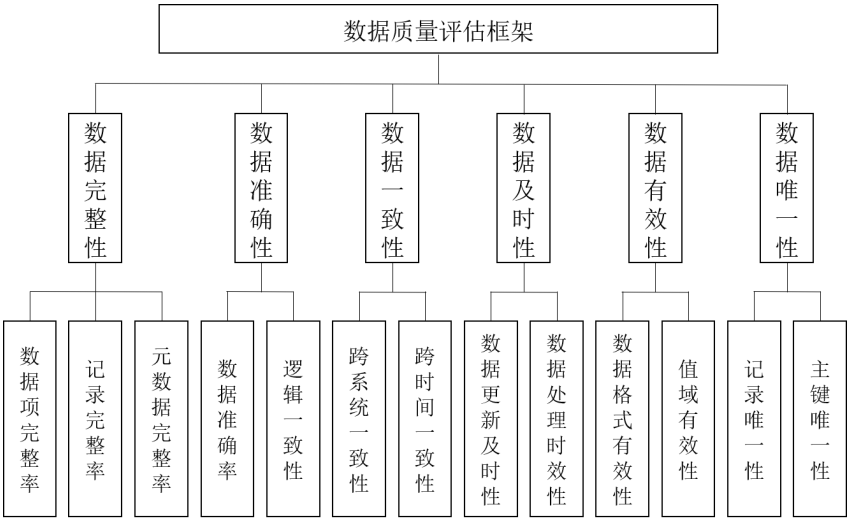


图 1 数据质量评估指标体系框架

5.2 评估指标及说明

围绕数据完整性、数据准确性、数据一致性、数据及时性、数据有效性、数据唯一性等维度，分别针对结构化数据、半结构化数据、非结构化数据构建数据质量评估指标。针对用于人工智能大模型训练的高质量数据集，应额外评估语料多样性、标注一致性、样本均衡性等专项指标。指标体系及其说明详见附录A。

6 评分计算方法

6.1 数据集质量评分模型

数据集质量评分用于反映单个数据集的数据质量水平，是部门整体数据质量评价的基础依据。本规范采用加权算术平均法计算数据集综合质量评分。以各基础质量指标权重作为加权系数，对各基础质量指标得分进行加权汇总，并除以权重总和，得到数据集综合质量评分。数据集综合质量评分按照公式（1）计算，基础质量指标得分按照公式（2）计算，检测规则得分按照公式（3）计算。

$$D = \frac{\sum_{i=1}^v (R_i \times W_i)}{\sum_{i=1}^v W_i} \dots\dots (1)$$

符号释义：

D ：数据集综合得分

R_i ：第 i 项基础质量指标得分

W_i ：第 i 项基础质量指标权重

v ：本次参评基础质量指标总数量（ $v \leq n$ ）

基础质量指标包括：

- 数据完整性
- 数据准确性
- 数据一致性
- 数据及时性
- 数据有效性
- 数据唯一性

$$R_i = \frac{\sum_{j=1}^u (S_{ij} \times Y_{ij})}{\sum_{j=1}^u Y_{ij}} \dots\dots (2)$$

符号释义：

j ——第 j 个实际启用的质量检测规则；

u ——第 i 个质量指标下实际启用的检测规则数量（ $u \leq m$ ）；

m ——第 i 个质量指标下的检测规则数量；

S_{ij} ——第 j 个检测规则的实际得分；

Y_{ij} ——第 j 个检测规则的权重。

$$S_{ij} = \frac{A_{ij}}{B_{ij}} \times 100 \dots\dots (3)$$

符号释义：

S_{ij} ——第 j 个检测规则得分；

A_{ij} ——符合检测规则的数据记录数量；

B_{ij} ——参与检测的数据记录总数。

注：未上报、空表或无有效数据的数据集，基础质量指标质量评分按 0 分计算；新接入且接入时间不足一个评估周期的数据集，可暂不纳入本周周期质量评分。

注：当分母（如总数据项数、总记录数、总文件数等）为零时，对应指标得分按 0 分计算。对于未设

置主键的数据集，应以业务唯一标识或组合字段作为唯一性校验依据；确无唯一标识的，应在元数据中注明原因并在评估报告中说明。各基础质量指标得分的计算应明确得分项与扣分项的判定标准，确保评分过程透明可追溯。

6.2 部门数据质量评分模型

部门质量评分用于综合评价数据提供单位整体数据质量水平，以纳入评估范围的数据集质量评分为基础进行汇总计算，最终得分按照公式（4）计算。

$$G = \frac{\sum_{k=1}^p (D_k \times W_k)}{\sum_{k=1}^p W_k} \dots\dots (4)$$

- 符号释义：
- G——数据提供单位质量最终得分；
 - D_k——第 k 个数据集质量评分；
 - W_k——第 k 个数据集权重；
 - p——纳入评价范围的数据集数量。

6.3 数据集权重配置

数据集指标权重应结合指标业务重要程度、数据体量、质量问题影响程度、数据更新频率等因素综合确定。其中，数据完整性、数据准确性等核心基础评价维度可适当提高权重；仅用于辅助校验、无强制业务管控要求的次要指标可降低权重。长期停用、仅作临时统计的辅助指标原则上不纳入综合得分权重计算范围。

6.4 评估等级划分

表 1 数据质量评估等级划分

评估等级	综合得分范围	质量描述
优秀	90-100 分	数据质量非常高，完全满足业务需求
良好	80-89 分	数据质量较高，基本满足业务需求
一般	70-79 分	数据质量中等，部分满足业务需求
较差	60-69 分	数据质量较低，难以完全满足业务需求
极差	0-59 分	数据质量很差，基本不能满足业务需求

7 数据质量评估程序

7.1 评估流程

数据质量评估流程见图2，依次为评估准备、指标选择、规则确定、评估实施、核实整改、数据应用、评估总结等步骤。评估应覆盖数据全生命周期，自数据产生阶段起纳入评估范围，确保评估的连续性和完整性。

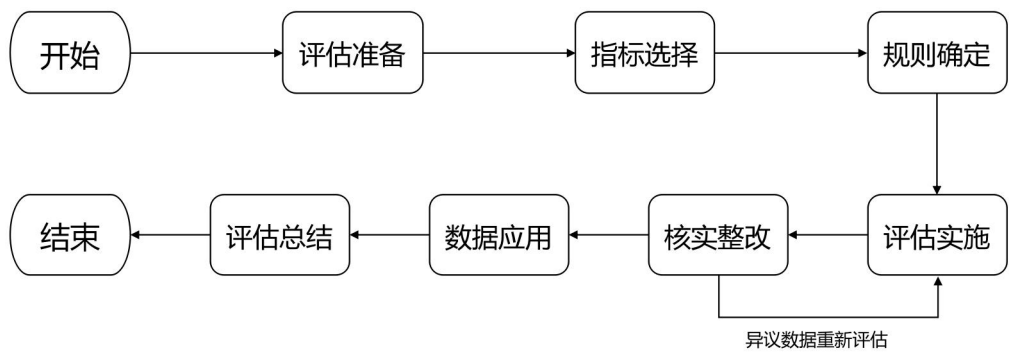


图 2 流程图

7.2 评估准备阶段

完成评估前期筹备相关工作，包括但不限于明确业务范围和数据、组建评估团队并划分分工、确定评估指标方法、形成实施方案、配置评估工具，输出配套文档等。

7.3 指标选择阶段

对照数据质量评估指标，结合业务需求筛选适配指标，按重要度分配指标权重分值，输出各阶段指标清单与配置表。

7.4 规则确定阶段

结合业务实际情况，制定数据量、数据项校验、跨表关联校验等规则，划分数据缺陷异常等级，通过数据测试修正逻辑，最终形成完整检核规则库。

7.5 评估实施阶段

采集待评估数据，计算各项质量指标，识别数据质量问题并梳理成因，依次输出指标计算结果、问题清单与分析报告。

7.6 核实整改阶段

结果检验阶段复核修正异议数据，核算数据集综合质量得分，定位各类数据问题根源，交由数源部门核实，非问题数据的情况须反馈说明，确实为问题数据的情况须及时整改，形成配套明细、统计台账与确认记录。整改后的异议数据返回上一阶段重新评估实施，异议数据处理相关要求参考《DB32/T 4608.5-2026 公共数据管理规范 第5部分：数据异议处理》。

7.7 数据应用阶段

质检通过的数据按照业务需要应用在具体业务场景，实现数据赋能效果。数据使用过程中发现的问题或者投诉，及时反馈至上一级数据提供部门（例如数据使用部门应把相关问题反馈至数据管理部门），由数据提供部门限时核实处理，异议数据处理相关要求参考《DB32/T 4608.5-2026 公共数据管理规范 第5部分：数据异议处理》。

7.8 评估总结阶段

评估总结阶段汇总全部评估结果，划定数据质量等级，编制正式评估报告并向相关责任方反馈留存反馈记录。

8 数据质量持续改进

构建常态化长效质量管理机制，循环迭代提升数据质量，实现数据质量长效稳定提升。

1. 数据提供方按照国家及我省数据治理、数据质量评估、异议处理等相关标准做好源头数据治理和质量评估，提供高质量数据；

2. 常态化开展数据质量评估，结合业务更新、治理需求动态优化质量校验规则、评估指标体系，及时反馈评估结果至数源部门；

3. 数据提供方按照《DB32/T 4608.5-2026 公共数据管理规范 第5部分：数据异议处理》要求做好数据异议核实处理，及时整改异议数据；

4. 数据使用方及时总结应用经验，反馈应用成效、异议数据等信息至数据管理部门；

5. 数据提供方、数据管理方及时总结经验、处理异议数据，支持好数据使用部门业务开展，多方协同持续优化“供数—评估—核实整改—应用—优化提升”的闭环质量管理模式，推动公共数据质量管理全流程制度化、标准化、规范化运行。

附录 A
(规范性)
数据质量指标详细说明

附录 A 数据质量指标说明表

表 A.1 数据质量指标详细说明表

一级指标	二级指标	数据类型	指标说明	计算方式	示例
数据完整性	数据项完整率	结构化数据	业务规范明确要求的所有必填字段无空值、无空白、无无效占位符，身份证号、统一社会信用代码、证件编号、主体名称等核心关键字段必须 100%填充完整，无缺失。 针对江苏省人口基础库、法人基础库等核心数据，行政区划代码应统一采用 GB/T 2260 标准，组织机构代码应采用统一社会信用代码（GB 32100-2015），性别代码应统一映射为国标 1-男、2-女、0-未知。 针对电子证照库数据，应额外评估证照编号唯一性、证照有效期标注完整性及证照状态（有效/过期/注销）标识准确率。针对社会信用库数据，应重点评估信用记录完整性、更新及时性及信用	（必填非空有效字段数 ÷ 全部必填字段总数）×100%	例如某企业信息表规定"统一社会信用代码""企业名称""法定代表人"为必填字段，若某条记录的"统一社会信用代码"为空或填写"N/A"，则判定该记录数据项不完整。

一级指标	二级指标	数据类型	指标说明	计算方式	示例
			评分口径的规范性。针对自然资源和空间地理库数据，应重点评估空间坐标精度、地理信息现势性及 GIS 数据格式的标准化程度。		
	记录完整率		按照既定采集周期、业务归集规则，应归集的全部业务数据记录无遗漏、无断档、无批次缺失，时序数据连续完整，无批量数据缺失情况。	$(\text{实际有效归集记录数} \div \text{业务应采集总记录数}) \times 100\%$	例如某部门每日应归集 1000 条业务记录，当天实际仅归集 980 条，则缺失 20 条记录，记录完整率为 98%，判定存在数据缺失。
	元数据完整率		结构化数据表、数据记录配套的元数据信息完整规范，数据来源、采集时间、更新周期、数据版本、业务归属、存储路径等关键属性无缺失。	$(\text{元数据完整有效条目数} \div \text{总数据条目数}) \times 100\%$	例如某张数据表缺少"数据来源"或"更新时间"等元数据属性，则判定该数据表元数据不完整。
	文件完整率	非结构化数据	图像、视频文件本体完整无损，无文件截断、半截损坏、编码异常、无法打开等问题，可正常预览、播放、解析，文件大小、格式参数完整合规。	$(\text{完整可正常使用文件数} \div \text{总文件数}) \times 100\%$	例如某段执法视频播放至一半即中断，或图片无法正常打开，则判定该文件不完整。
	元数据完整率		图像、视频配套元数据齐全，包含采集时间、采集设备、采集地点、分辨率、时长、业务场景、操作人员等关键信息，无关键属性空白。	$(\text{元数据完整文件数} \div \text{总文件数}) \times 100\%$	例如某张现场照片缺少采集时间和采集地点信息，则判定该文件元数据不完整。

一级指标	二级指标	数据类型	指标说明	计算方式	示例
	音轨完整性		带音频的视频音轨完整有效，无无声、音轨丢失、杂音、断音、音频失真等问题，音轨与视频文件绑定完整。	$(\text{音轨完整有效视频数} \div \text{总带音频视频数}) \times 100\%$	例如某视频画面正常但全程无声音，或音频存在严重杂音、断续现象，则判定音轨不完整。
数据准确性	数据准确率	结构化数据	数据内容真实有效，与权威公示数据、客观实体信息、国家标准内容完全一致，无虚假填报、错误录入、不实修改等问题，精准反映业务实际情况。应通过跨系统比对，确保同一政务主体在源系统与"苏政源"底座中的信息保持一致。	$(\text{内容完全正确数据项数} \div \text{总抽检/全量数据项数}) \times 100\%$	例如企业统一社会信用代码与国家企业信用信息公示系统登记信息不一致，或身份证号码录入错误，则判定数据内容不准确。
	逻辑准确性		单条数据内部、关联数据之间符合通用常识和专属业务逻辑，无逻辑冲突、常识错误、因果矛盾，适配业务运行规则。	$(\text{逻辑合规数据项数} \div \text{总数据项数}) \times 100\%$	例如人员出生日期晚于参加工作日期，或注销企业状态仍显示正常经营，则判定存在逻辑错误。
	内容真实性	非结构化数据	图像、视频为现场原始采集素材，无后期拼接、PS 修改、恶意剪辑、画面替换、篡改伪造等行为，完整保留原始采集状态，真实还原业务场景。	工具检测+人工抽检 核验	例如某现场执法照片经检测发现存在后期拼接或 AI 生成痕迹，或视频存在恶意剪辑、关键画面删除等情况，则判定内容真实性不符合要求。
	场景匹配率		文件内容与采集标注的业务场景、事件时间、地点、业务类型完全匹配，无错拍、误拍、场景不符、事件错位	$(\text{场景匹配合规文件数} \div \text{抽检总文件数}) \times 100\%$	例如标注为"垃圾分类执法现场"的视频实际拍摄内容为道路巡查，或采集地点与标注地点不一致，则判定场景不匹配。

一级指标	二级指标	数据类型	指标说明	计算方式	示例
			问题。		
数据一致性	跨系统一致性	结构化数据	同一政务主体、同一业务数据，在多套业务系统、数据库、共享平台中存储的信息保持一致，数据变更可同步更新。特别关注江苏省"一网通办"、"长三角跨省通办"等场景下，人口基础库、法人基础库等核心数据在公安、市场监管、社保等多部门系统间的一致性。 针对自然资源和空间地理库数据，应重点评估地理空间数据在不同部门系统间的坐标系一致性、更新版本同步及地图图层数据的全面性与及时性。	$(\text{跨系统比对一致数据项数} \div \text{总比对数据项数}) \times 100\%$	例如某企业在市场监管系统中的企业名称已变更为"XX 科技有限公司"，但共享交换平台仍显示旧名称，则判定跨系统数据不一致。
	跨时间一致性		同一数据的历史版本与现行版本无无故篡改、无前后矛盾，静态基础数据保持稳定，动态数据变更留痕、合规可追溯。	$(\text{跨时间比对一致数据项数} \div \text{总比对数据项数}) \times 100\%$	例如某企业统一社会信用代码在历史版本与当前版本中无业务依据发生变化，且无变更记录，则判定跨时间数据不一致。
	参数一致性	非结构化数据	同一设备、同一批次、同一场景采集的图像分辨率、画质参数、视频帧率、码率、编码格式保持一致，无参数参差不齐、规格混乱问题。	$(\text{参数一致合规文件数} \div \text{同批次总文件数}) \times 100\%$	例如同一批次采集的视频要求分辨率为1920×1080、帧率 30fps，但部分视频为 1280×720 或 15fps，则判定参数不一致。

一级指标	二级指标	数据类型	指标说明	计算方式	示例
	嵌套层级一致性		半结构化数据中数组嵌套深度不一。基于 JSONPath/XPath 路径进行深层遍历校验。	$(\text{嵌套层级一致数} \div \text{同批总数}) \times 100\%$	例如同一个接口返回的 JSON 数据中，正常结构为 "user.address.city"，部分数据缺少 address 节点或增加多层嵌套，导致解析失败，则判定嵌套层级不一致。
	内容关联一致性		同一政务事件对应的多张图像、多段视频素材，场景主体、事件流程、画面内容连贯统一，无内容冲突、逻辑断裂问题。	人工抽检定性判定	例如同一次执法检查的照片显示检查对象为 A 企业，而对应视频记录为 B 企业，或事件时间前后矛盾，则判定内容关联不一致。
数据及时性	数据更新及时性	结构化数据	严格按照实时、每日、每周、每月等既定更新周期完成数据增量更新、全量刷新，动态业务数据无超期未更新、长期滞留情况。针对具有生命周期的数据（如临时通行证、短期政策文件等），应评估过期数据的自动标记或归档情况，避免过期数据误导业务决策。 针对社会信用库数据，应重点评估信用记录的更新周期是否满足业务需求，以及信用状态变更的同步及时性； 针对自然资源和空间地理库数据，应重点评估地理信息的现势性及更新周期是否满足地图更新要求。	$(\text{按时完成更新数据项数} \div \text{应更新数据项总数}) \times 100\%$	例如人口基础信息要求每日更新，但某条记录已超过 3 天未同步更新，则判定数据更新不及时。

一级指标	二级指标	数据类型	指标说明	计算方式	示例
	数据处理时效性		数据可在规定时限内完成清洗、校验、入库、归档处理，无数据堆积、处理超时、积压滞留情况，保障数据流转高效。	$(按时完成处理数据量 \div 总待处理数据量) \times 100\%$	例如数据治理平台规定新增数据 2 小时内完成清洗入库，若某批数据处理耗时超过 2 小时，则判定处理时效不符合要求。
	采集及时性	非结构化数据	事件发生、业务开展过程中，按规范及时完成图像抓拍、视频录制，无事件结束后滞后采集、漏采、延时留存问题。	$(及时采集合规文件数 \div 总文件数) \times 100\%$	例如道路交通事故发生时间为 10:00，现场照片于当天 15:00 补拍，无法真实反映现场情况，则判定采集不及时。
	转码/归档及时性		原始视频、图像素材在规定时限内完成标准化转码、切片、压缩处理，按期完成归档入库，无素材堆积、归档超时问题。	定性合规判定	例如视频采集完成后要求 24 小时内完成转码并归档，但实际超过 48 小时仍未完成处理，则判定转码或归档不及时。
数据有效性	数据格式有效性	结构化数据	身份证号、手机号、统一社会信用代码、日期、邮编、证件编号等字段格式严格符合国家标准规范，数据长度、编码规则无偏差。时间字段应统一转换为 YYYY-MM-DD HH:mm:ss 格式，金额字段应统一保留两位小数并明确币种。	$(格式完全合规数据项数 \div 总数据项数) \times 100\%$	例如身份证号码应为 18 位且符合校验规则，若录入为 15 位、长度错误或包含非法字符，则判定数据格式无效。
	值域有效性		所有数据取值严格限定在国标代码、行业规范、业务预设值域范围内，如性别、行政区划、业务状态等字段无	$(值域合规数据项数 \div 总数据项数) \times 100\%$	例如"性别"字段规定取值为"男""女"或代码"1""2"，若出现"未知""3""A"等非法取值，则判定值域无效。

一级指标	二级指标	数据类型	指标说明	计算方式	示例
			越界取值、无效代码、非法内容。		
	命名有效性	非结构化数据	严格遵循政务非结构化数据命名规范，文件名包含时间、场景、设备、业务类型等关键信息，无乱码、特殊字符、无规则命名，便于统一管理检索。	$(\text{命名合规文件数} \div \text{总文件数}) \times 100\%$	例如文件命名规范要求为"20260714_执法检查_设备 001.jpg"，若文件名为"IMG001.jpg""新建文件.jpg"或包含乱码字符，则判定命名不规范。
	文件格式有效性		图像、视频采用政务通用标准格式，图像支持 JPG、PNG、BMP 等，视频支持 MP4、AVI 等通用格式，可正常解析、预览、流转。	$(\text{格式合规文件数} \div \text{总文件数}) \times 100\%$	例如系统规定上传 MP4 视频，若上传 RMVB、FLV 或无法识别的格式，导致平台无法正常播放，则判定文件格式无效。
数据有效性	记录唯一性		同一数据集内无完全重复、高度相似的业务记录，无批量冗余数据，每条业务记录对应唯一业务场景，无重复归集问题。	$(\text{唯一不重复记录数} \div \text{数据集总记录数}) \times 100\%$	例如同一企业、同一办理事项、同一受理时间的数据重复归集两次，则判定存在重复记录，不符合记录唯一性要求。
	主键唯一性		数据表主键字段全局唯一、无重复、无空值，作为数据唯一标识精准有效，是数据关联、检索、更新的核心保障，实行零容忍标准。	$(\text{唯一有效主键数} \div \text{总主键数}) \times 100\%$	例如数据表以"业务流水号"作为主键，若存在两个记录使用相同流水号，或主键为空，则判定主键唯一性不符合要求。
	文件唯一率	非结构化数据	素材库内无完全重复、同源多副本的图像、视频文件，无重复上传、冗余存储问题，每条素材对应唯一采集记录。	$(\text{唯一无重复文件数} \div \text{总文件数}) \times 100\%$	例如同一张现场照片因重复上传保存了多个完全相同的文件，或同一视频存在多个重复副本，则判定文件唯一率不符合要求。

一级指标	二级指标	数据类型	指标说明	计算方式	示例
	隐私脱敏合规率	结构化数据	数据集中敏感个人信息（PII）、商业秘密等是否已按规范完成脱敏处理，确保对外共享、测试环境使用中的数据符合隐私保护要求。	（通过隐私检测工具扫描无敏感信息的记录数 / 总记录数）×100%	例如某数据集包含公民身份证号、手机号等敏感字段，对外共享前未进行脱敏处理，则判定隐私脱敏合规率不符合要求。
	标注一致性	结构化数据	针对用于人工智能大模型训练的数据集，评估不同标注员之间标注结果的一致性水平，确保标注质量满足模型训练需求。	$Kappa = (Po - Pe) / (1 - Pe)$ ，其中 Po 为观察到的一致性比例，Pe 为随机期望一致性比例	例如某政务文本分类标注任务中，两名标注员对同一批数据的分类结果 Kappa 系数低于 0.6，则判定标注一致性不达标，需重新统一标注规范。

参 考 文 献

- [1] GB/T 34960.5—2018 信息技术服务 治理 第5部分：数据治理规范
- [2] GB/T 36344—2018 信息技术数据质量评价指标
- [3] GB/T 36073—2018 数据管理能力成熟度评估模型
- [4] GB/T 44109—2024 信息技术 大数据 数据治理实施指南
- [5] 张彬编著. 数据治理理论、方法与实践[M]. 清华大学出版社, 2024
- [6] 梅宏主编. 数据治理之法[M]. 中国人民大学出版社, 2022