ICS 35.240.01
CCS L 70

团 体 标 准

T/HAAI 003—2024

数据资产 数据质量评价规范

Data assets— Data quality evaluation specification

2024－11－27 发布 2024－12－01 实施

海南省人工智能学会 发 布

目　　次

前　言

本文件按照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》的规定起草。

本文件由海南省人工智能学会提出并归口。

本文件起草单位：中国电信股份有限公司数据要素技术创新（海南）中心、中国电信股份有限公司海南分公司、海南师范大学、武汉大学、海南科技职业大学。

本文件主要起草人：张小建、黄健强、陈文思、符舒凡、梁钰、吴佩琦、丁超、黄程杰、周政成、郭世坤、蔡雪云、严炜炜、王艺臻、白颢。

数据资产 数据质量评价

1 范围

本文件规定了数据资产质量评价的评价内容、评价指标和评价流程的应用。

本文件适用于所有公共和私营部门的数据资产质量评价工作。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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 36344-2018 《信息技术 数据质量评价指标》

海南省数据资源标准规范 第7部分:数据标准与质量管理规范

其他相关地区和行业标准

3 术语和定义

GB/T 36344-2018界定的以及下列术语和定义适用于本文件。

3.1

数据资源 data resources

是指具有使用价值的数据，是可供人类利用的新型资源。

3.2

数据产品 data products

是指基于数据加工形成的，可满足特定需求的数据加工品和数据服务。

3.3

数据资产 data assets

是指特定主体合法拥有或者控制的，能进行货币计量的，且能带来直接或者间接经济利益的数据资源。

[来源:GB/T 40685-2021，3.1，有修改]

3.4

数据质量 data quality

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 36344-2018，2.3]

3.5

数据质量评价 data quality evaluation

按照数据质量评价指标体系，采用适当的方法对数据质量进行评估，并形成数据质量评价结果的过程。

[来源：福建省地方标准 DB35/T 1952-2020,3.5]

注：数据质量评价内容包含数据的规范性、完整性、准确性、一致性、时效性、可访问性。

4 数据质量评价的目的与作用

本标准所述的数据质量评价是在完成数据治理、数据资产识别与分类、确权登记后进行的步骤，数据质量评价可为后续的数据资产入表、数据资产评估提供有力支撑，为数据交易标的的流通定价提供一定依据，有利于优化数据管理、实现数据资产保值增值。

数据质量评价旨在确保数据的规范性、完整性、准确性、一致性、时效性、可访问性，以支持相关业务决策与运营。

5 数据质量评价内容

5.1 评价内容

分类本标准按照以下指标框架的六个维度对数据质量进行分析定义，如图 1 所示。

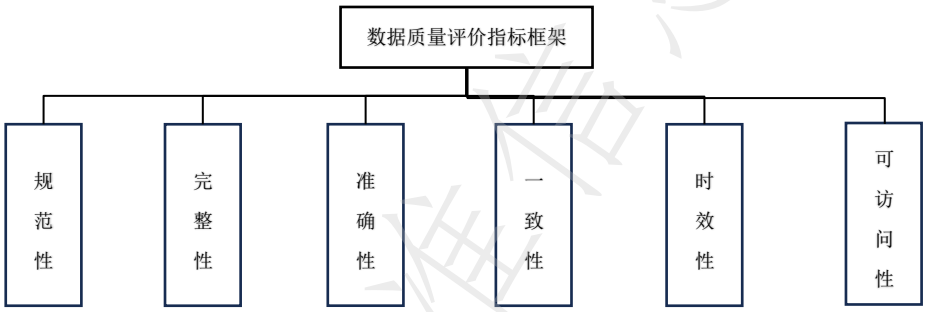

图 1 数据质量评价指标框架

具体规定如下：

——规范性：数据符合数据标准、数据模型、元数据、业务规则、权威参考数据或安全规范的度量；

——完整性：包括数据元素的完整性和数据记录的完整性；

——准确性：包括数据内容的正确性、数据重复率、数据唯一性和脏数据出现率；

——一致性：包括相同数据一致性和关联数据一致性；

——时效性：包括基于时间段的正确性、基于时间点的及时性和时序性；

——可访问性：数据在需要时可以获取，在设定的有效生存周期内可以使用。

5.2 评价方法

数据质量评价按照评价指标对数据质量进行数据探查、数据检核，数据质量评价方法一般使用定量评价法，其中：

——数据探查，是指探索源数据的过程，用来理解数据结构、数据内容、数据关系以及为数据工程识别可能存在的问题。在进行数据质量评价前，需要对数据进行探查，以确定各项数据的质量情况，为数据质量评价提供明细。

——数据检核一般使用定量评价法，包括但不限于以下方法：

- 全量检核：对涉及的所有数据进行逐一检核数据质量。
- 增量检核：对涉及的数据，在特定的范围和时间段内新增的数据进行逐一检核数据质量。
- 抽样检核：按照一定的抽样规则，对涉及的数据样本进行逐一检核数据质量。

6 数据质量评价指标

根据数据质量评价的六个维度，本标准按照各指标的使用频率以及范围对数据质量评价指标体系的一级指标分数进行界定，二级指标分数由评价规则进行具体界定。

6.1 规范性

规范性评价指标详见表1。

表1 规范性评价指标

序号	一级指标	一级指标分数	二级指标	二级指标分数	指标描述	评价方法	计算公式
1	规范性	X_1	数据标准规范检查	X_{11}	通过检查特定字段数据是否符合数据标准	经验分析、词组对比分析	$X_{1i}=\frac{W_{1i}}{N}\times X_1$ $N=\sum_{i=1}^{2}(W_{1i})$ 式中： W_{1i}=本二级指标对应的规则条数； N=本一级指标规则条数
2			行业参考数据规范检查	X_{12}	通过检查特定字段数据是否符合业务或行业主管部门相关参考数据及规则	经验分析	

6.2 完整性

完整性评价指标详见表2。

表2 完整性评价指标

序号	一级指标	一级指标分数	二级指标	二级指标分数	指标描述	评价方法	计算公式
1	完整性	X_2	字段信息缺失检查	X_{21}	实体关键属性是否存在空值	缺失值分析	$X_{2i}=\frac{W_{2i}}{N}\times X_2$ $N=\sum_{i=1}^{3}(W_{2i})$ 式中： W_{2i}=本二级指标对应的规则条数； N=本一级指标规则条数
2			数据记录检查	X_{22}	数据记录信息、数据链条是否存在缺失	缺失值分析	
3			代码值检查	X_{23}	表字段属性的代码值是否和数据字典一致	逻辑分析、经验分析	

6.3 时效性

时效性评价指标详见表3。

表3　时效性评价指标

序号	一级指标	一级指标分数	二级指标	二级指标分数	指标描述	评价方法	计算公式
1	时效性	X_3	更新及时性检查	X_{31}	基于时间戳的记录数、频率分布或延迟时间符合业务需求的程度，即数据实时更新的及时性	逻辑分析、经验分析	$X_{3i}=\frac{W_{3i}}{N}\times X_3$ $N=\sum_{i=1}^{1}(W_{3i})$ 式中： W_{3i}=本二级指标对应的规则条数； N=本一级指标规则条数

6.4　准确性

准确性评价指标详见表4。

表4　准确性评价指标

序号	一级指标	一级指标分数	二级指标	二级指标分数	指标描述	评价方法	计算公式
1	准确性	X_4	脏数据检查	X_{41}	检查特定字段的取值是否在预定的数字取值范围之内，或是否存在乱码、特殊符号等不可识别值，即检查无效数据的度量	经验分析	$X_{4i}=\frac{W_{4i}}{N}\times X_4$ $N=\sum_{i=1}^{3}(W_{4i})$ 式中： W_{4i}=本二级指标对应的规则条数； N=本一级指标规则条数
2			数据唯一性检查	X_{42}	检查特定字段、记录、文件或数据集唯一性的度量	重复值分析	
3			数据正确性检查	X_{43}	检查表字段的数据类型、长度、值域、精度等取值是否符合标准	经验分析	

6.5　一致性

一致性评价指标详见表5。

表5　一致性评价指标

序号	一级指标	一级指标分数	二级指标	二级指标分数	指标描述	评价方法	计算公式
1	一致性	X_5	存在一致性检查	X_{51}	检查同一数据在不同位置或被不同用户使用时是否保持一致	逻辑关系分析、经验分析	$X_{5i}=\frac{W_{5i}}{N}\times X_5$ $N=\sum_{i=1}^{2}(W_{5i})$ 式中： W_{5i}=本二级指标对应的规则条数； N=本一级指标规则条数
2			关联性检查	X_{52}	检查数据发生变化时，存储在不同位置的同一数据是否被同步修改	逻辑关系分析、经验分析	

6.6 可访问性

可访问性评价指标详见表6。

表6 可访问性评价指标

序号	一级指标	二级指标	指标描述	备注
1	可访问性	可访问检查	数据在需要时的可获取性	探查数据默认为可完整访问
2		实时访问中断检查	实时数据访问中断程度的度量	

6.7 评价指标参考值

根据政务数据的质量评价经验，本标准对数据质量评价指标提供参考值，分数如表7所示。一级指标具体分数根据指标设计及规则制定而变化，行业属性将对其产生影响。

表7 评价指标参考值

序号	指标	分数
1	规范性	X_1=15
2	完整性	X_2=30
3	时效性	X_3=10
4	准确性	X_4=30
5	一致性	X_5=15
合计		100

7　数据质量评价流程

7.1　总体流程

数据质量评价流程包括准备阶段、实施评价、评价结果、评价反馈，评价结果得出需要提升数据质量的，将根据要求进行反馈，更新后再次实施评价。具体流程如图2。

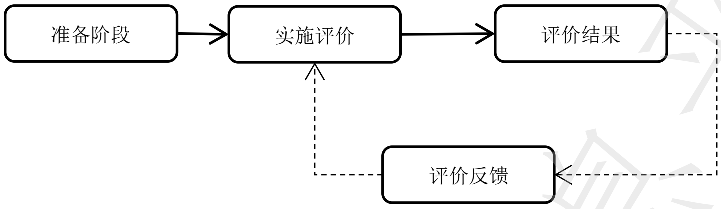

图2 数据质量评价总体流程

7.2　准备阶段

数据质量评价的准备阶段相关任务包括四个步骤：

——需求提出。数据质量评价需求方根据评估目的、数据情况提出评估需求。

——指标设计。根据数据指标体系选定具体评价指标，原则上要求至少包含规范性、完整性和准确性三项基础指标。本阶段需根据指标制定对应的规则，各二级指标对应的规则条数为W、一级指标规则条数为N，规则条数将影响各指标具体分数。

——措施制定。实施评价的措施包括但不限于审查文件和记录、观察数据质量管理过程和活动、人员访谈、工具测评、配置检查、旁站式验证等。

——数据准备。评价需求方按照专业评价机构要求，采用全量、总体抽样、分组抽样等方式，按照一定的比例或条件，将抽样的数据作为被待评价数据。

——确定分数计算。

•通过公式（1）计算二级指标分数：

$$X_{ki} = \frac{W_{ki}}{N} \times X_k \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（1）$$

式中：

k ——第 k 个一级指标；

i ——第 i 个二级指标；

X_{ki} ——二级指标分数；

W_{ki} ——二级指标规则条数；

N ——一级指标规则条数；

X_k ——一级指标分数。

二级指标分数之和为一级指标分数，且一级指标分数之和为100。具体分值分配将根据相应规则进行增减调整。

•通过公式（2）计算一级指标规则条数：

$$N = \sum_{i=1}^{j} (W_{ki}) \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（2）$$

式中：

k ——第 k 个一级指标；

i ——第 i 个二级指标；

j ——二级指标总数；

N ——一级指标规则条数；

W_{ki} ——二级指标规则条数。

7.3 实施评价

按照准备阶段制定的数据评价指标，对待评价数据集进行评价和记录。具体包括：

——专业评价机构在保证数据安全的条件下获取评价数据，并核实评价数据的规范性、完整性、时效性、准确性、一致性以及可访问性，记录评价的客观证据并根据已制定的规则与分数对数据进行评分。

通过公式（3）计算二级指标得分：

$$T_2=\frac{A}{B}\times X_{ki} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{（3）}$$

式中：

T_2——二级指标得分；

A——符合指标要求的数据量；

B——指标对应总数据量；

X_{ki}——二级指标分数；

k ——第 k 个一级指标；

i ——第 i 个二级指标。

——对于专业评价机构不能确认的数据，可采取再检查核对的方式给予确认。评价机构与评价需求方要对数据评价中提出的问题进行验证，并做出相关记录。

——专业评价机构应综合评价实施过程中的各项结果，整理评价需求方发现的与数据质量评价数据不符合的问题。

——对于待确认的数据，评价需求方应按照专业评价机构的需求提供的证据进行再次审查，评价机构验证新证据是否有效，若新证据有效，则基于新证据调整评价结果。

7.4 评价结果

专业评机构按照数据质量评价指标进行具体评分，再根据评分结果对数据质量进行定级。分数与评价等级划分如表8所示。

表8 评价等级区间表

评价指标	分数区间	区间	评价等级
表级评价	80 分及以上	[80,100]	好
	60 分（含）至 80 分	[60,80)	一般
	60 分以下	(60,0]	差

具体分数计算公式如下：

结合检查结果与数据质量评价标准、计算公式算出数据质量得分。计算步骤如下所示：

先通过公式（4）计算二级指标的得分：

$$T_2=\frac{A}{B}\times X_{ki} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{（4）}$$

式中：

T_2 ——二级指标得分；

A ——符合指标要求的数据量；

B ——指标对应总数据量；

X_{ki} ——二级指标分数；

k ——第 k 个一级指标；

i ——第 i 个二级指标。

在（4）的基础上，通过公式（5）计算一级指标的得分：

$$T_1=\sum_{i=1}^{n} T_2=(T_{21}+T_{22}+\cdots+T_{2N}) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（5）$$

式中：

T_1 ——一级指标得分；

T_2 ——二级指标得分；

T_{21} ——各项二级指标得分。

在（5）的基础上，通过公式（6）计算表的得分：

$$T=\sum_{i=1}^{n} T_1=(T_{11}+T_{12}+\cdots+T_{1N}) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（6）$$

式中：

T_1 ——一级指标得分；

T_{11} ——各项一级指标得分。

7.5 评价反馈

评价反馈阶段包括三个步骤：

——分析评价结果：首先分析评价结果中发现的问题，包括规范性、完整性、时效性、准确性、一致性、可访问性六个维度，根据问题的严重程度、对数据使用的影响程度以及解决问题的难易程度，对识别出的问题进行优先级排序。

——制定与实施改进措施：根据每个问题制定并实施相应的改进措施，包括技术、管理等层面的优化。

——沟通与培训：确保相关人员都了解当前数据质量情况以及相应改进措施，适时举办数据质量相关培训以提升相关人员的数据质量意识与技能水平。

参 考 文 献

[1] GB/T 7027-2002 信息分类和编码的基本原则与方法
[2] GB/T 37550-2019 电子商务数据资产评价指标体系
[3] GB/T 38673-2020 信息技术 大数据 大数据系统基本要求
[4] GB/T 42450-2023 信息技术 大数据 数据资源规划
[5] GB/T 36344-2018 信息技术 数据质量评价指标
[6] GB/T 40685-2021 信息技术服务 数据资产 管理要求
[7] DB35/T 1952-2020 福建省地方标准 数据质量评价规范 公共信息资源开放
[8] 海南省数据产品超市数据产品确权登记实施细则（暂行）,琼数运〔2023〕52 号
[9] 数据资产评估指导意见，中评协〔2023〕17号
[10]海南省数据资源标准规范 第7部分:数据标准与质量管理规范

附 录 A

数据质量评价结果明细表模板

表 A.1 数据质量评价结果明细表模板

数源单位	数据目录	表名	一级指标	二级指标	分数	二级指标得分	一级指标得分	表级得分	备注