ICS 35.240.01
L 70

中华人民共和国国家标准

GB/T 36344—2018

信息技术　数据质量评价指标

Information technology—Evaluation indicators for data quality

2018-06-07 发布　　2019-01-01 实施

国家市场监督管理总局
中国国家标准化管理委员会　发布

目　　次

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

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这些专利的责任。

本标准由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本标准起草单位:中国电子技术标准化研究院、御数坊(北京)科技咨询有限公司、上海市信息投资股份有限公司、中国科学院计算机网络信息中心、深圳市华傲数据技术有限公司、贵阳信息技术研究院(中科院软件所贵阳分部)、国网浙江省电力有限公司。

本标准主要起草人:卫凤林、宾军志、甘似禹、胡良霖、于文渊、黎俊茂、陈峰、杨达、王静、董建、张群、张展新、赵菁华、李冰、李易昂、秦俊宁、陈利跃。

信息技术　数据质量评价指标

1　范围

本标准规定了数据质量评价指标的框架和说明。

本标准适用于数据生存周期各个阶段的数据质量评价。

2　术语和定义

下列术语和定义适用于本文件。

2.1

数据　data

信息的可再解释的形式化表示，以适用于通信、解释或处理。

注：可以通过人工或自动手段处理数据。

[GB/T 5271.1—2000，定义 01.01.02]

2.2

元数据　metadata

关于数据或数据元素的数据(可能包括其数据描述)，以及关于数据拥有权、存取路径、访问权和数据易变性的数据。

[GB/T 5271.17—2010，定义 17.06.05]

2.3

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。

2.4

原始数据　raw data

终端用户所存储使用的各种未经过处理或简化的数据。

注：原始数据有多种存在形式，如文本数据，图像数据，音频数据或者几种数据混合存在。

2.5

数据生存周期　data lifecycle

将原始数据转化为可用于行动的知识的一组过程。

2.6

数据集　dataset

具有一定主题，可以标识并可以被计算机化处理的数据集合。

2.7

数据模型　data model

对分析的图像和文本表述，该分析识别了组织为完成其使命、功能、目标、目的和战略，以及管理和评价组织所需要的数据。

注 1：在从高到低的不同抽象层次表示数据时，通常会区分概念模型(与某些努力相关的概念组成的模型)、逻辑模型和物理模型。

注 2：所使用数据模型的使用周境的边界的正规描述，称为上下文模式。

注 3：数据模型标识实体、域(属性)以及与其他数据的关系(关联)，提供数据和数据间关系的概念视图。

示例 1：由框图组成的语义数据模型，这种框代表对业务有意义的事务集，如“人”或“行动”，以及描述这类实体对之间关系的线条。

示例 2：应用特定数据管理技术的关系表或可扩展标记语言 XML 等是逻辑数据模型。

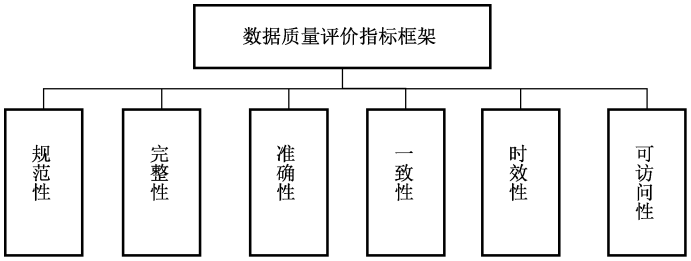
2.8

数据标准　data standard

数据的命名、定义、结构和取值规范方面的规则和基准。

3　指标框架

数据质量评价指标框架见图 1。



说明：

规范性——数据符合数据标准、数据模型、业务规则、元数据或权威参考数据的程度。

完整性——按照数据规则要求，数据元素被赋予数值的程度。

准确性——数据准确表示其所描述的真实实体(实际对象)真实值的程度。

一致性——数据与其他特定上下文中使用的数据无矛盾的程度。

时效性——数据在时间变化中的正确程度。

可访问性——数据能被访问的程度。

图 1　数据质量评价指标框架

4　概述

第 5 章规定的六大类评价指标，是实施数据质量评价的最小集，数据质量评价过程参见附录 A。

5　指标说明

5.1　评价表中表头信息说明

评价表中的表头说明如下：

a)　指标编号及编码规则：指标编号是评价指标的唯一性编号，由一级指标和二级指标共 4 位数字组成。编码规则见图 2。

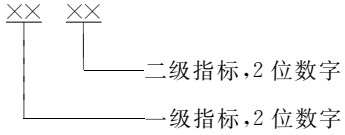


图 2　编码规则

1)　一级指标：由 2 位数字组成，01 代表规范性、02 代表完整性、03 代表准确性、04 代表一致

性、05 代表时效性、06 代表可访问性；

2） 二级指标：由 2 位数字组成的顺序码，范围为 01～99。

b） 指标名称：评价指标的名称。

c） 指标描述：评价指标的解释。

d） 计算方法：评价指标的计算方法。

5.2 规范性

规范性评价指标定义见表 1。

表 1 规范性评价指标

指标编号	指标名称	指标描述	计算方法
0101	数据标准	数据符合数据标准的度量。 **注 1**：评价数据质量时需要收集数据在命名、创建、定义、更新和归档时遵循的标准，包括国际标准、国家标准、行业标准、地方标准或相关规定等。 **注 2**：和数据归档一样甚至更重要，在一个完整的数据规则中旧数据的销毁一般也有一个比较详细且具有可执行性的规定	$X=A/B$ 式中： A＝满足数据标准要求的数据集中元素的个数； B＝被评价的数据集中元素的个数
0102	数据模型	数据符合数据模型的度量。 **注 1**：数据模型是一种直观描述组织数据结构的手段，是数据表达的规范。 **注 2**：评价数据质量时需要检查是否存在清晰可理解的数据模型定义以及这些数据的组织形式	$X=A/B$ 式中： A＝满足数据模型要求的数据集中元素的个数； B＝被评价的数据集中元素的个数
0103	元数据	数据符合元数据定义的度量。 **注**：元数据标注、描述或刻画其他数据、以使检索、或使用信息更容易。评价数据质量时需要检查是否提供可解读的元数据文档。 **示例**：包含各字段名称、描述、类型值域等内容的数据字典为一种元数据文档	$X=A/B$ 式中： A＝满足元数据定义的数据集中元素的个数； B＝被评价的数据集中元素的个数
0104	业务规则	数据符合业务规则的度量。 **注 1**：业务规则是一种权威性原则或指导方针，用来描述业务交互，并建立行动和数据行为结果及完整性的规则。 **注 2**：评价数据质量时需要检查是否存在良好归档的业务规则	$X=A/B$ 式中： A＝满足业务规则的数据集中元素的个数； B＝被评价的数据集中元素的个数
0105	权威参考数据（权威参考源）	参考数据是系统、应用软件、数据库、流程、报告及交易记录和主记录用来参考的数值集合或分类表。 **注**：评价数据质量时需要收集参考数据列表。 **示例**：一张用于一个特定字段的有效值列表为一种参考数据类型	$X=A/B$ 式中： A＝满足参考数据规则的数据集中元素的个数； B＝被评价的数据集中元素的个数
0106	安全规范	安全规范是安全和隐私方面的规则，包括数据权限管理，数据脱敏处理等	$X=A/B$ 式中： A＝满足安全规范的数据集中元素的个数； B＝被评价的数据集中元素的个数

5.3 完整性

完整性评价指标定义见表2。

表2 完整性评价指标

指标编号	指标名称	指标描述	计算方法
0201	数据元素完整性	按照业务规则要求，数据集中应被赋值的数据元素的赋值程度	$X=A/B$ 式中： A＝被赋值的数据集中元素的个数； B＝预期被赋值的数据集中元素的个数
0202	数据记录完整性	按照业务规则要求，数据集中应被赋值的数据记录的赋值程度	$X=A/B$ 式中： A＝被赋值的数据集中元素的个数； B＝预期被赋值的数据集中元素的个数

5.4 准确性

准确性评价指标定义见表3。

表3 准确性评价指标

指标编号	指标名称	指标描述	计算方法
0301	数据内容正确性	数据内容是否是预期数据	$X=A/B$ 式中： A＝满足数据正确性要求的数据集中元素的个数； B＝被评价的数据集中元素的个数
0302	数据格式合规性	数据格式（包括数据类型、数值范围、数据长度、精度等）是否满足预期要求。 **示例**：性别一栏不能出现男/女以外的内容；身份证号不能出现标点符号；以及对字符编码的一些限制，都需要通过规定内容的格式来实现	$X=A/B$ 式中： A＝满足格式要求的数据集中元素的个数； B＝被评价的数据集中元素的个数
0303	数据重复率	特定字段、记录、文件或数据集意外重复的度量	$X=A/B$ 式中： A＝重复的数据集中元素的个数； B＝被评价的数据集中元素的个数
0304	数据唯一性	特定字段、记录、文件或数据集唯一性的度量	$X=A/B$ 式中： A＝满足唯一性要求的数据集中元素的个数； B＝被评价的数据集中元素的个数
0305	脏数据出现率	正确字段、记录、文件或数据集之外无效数据的度量。 **示例**：事务发生回滚时由于回滚机制不健全或不完善导致可能出现脏数据	$X=A/B$ 式中： A＝有脏数据出现的数据集中元素的个数； B＝被评价的数据集中元素的个数

5.5 一致性

一致性评价指标定义见表4。

表 4　一致性评价指标

指标编号	指标名称	指标描述	计算方法
0401	相同数据一致性	同一数据在不同位置存储或被不同应用或用户使用时，数据的一致性；数据发生变化时，存储在不同位置的同一数据被同步修改	$X=A/B$ 式中： A＝满足一致性要求的数据集中元素的个数； B＝被评价的数据集中元素的个数
0402	关联数据一致性	根据一致性约束规则检查关联数据的一致性	$X=A/B$ 式中： A＝满足一致性要求的数据集中元素的个数； B＝被评价的数据集中元素的个数

5.6　时效性

时效性评价指标定义见表 5。

表 5　时效性评价指标

指标编号	指标名称	指标描述	计算方法
0501	基于时间段的正确性	基于日期范围的记录数或频率分布符合业务需求的程度	$X=A/B$ 式中： A＝满足有效性要求的数据集中元素的个数； B＝被评价的数据集中元素的个数
0502	基于时间点及时性	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	$X=A/B$ 式中： A＝满足及时性要求的数据集中元素的个数； B＝被评价的数据集中元素的个数
0503	时序性	数据集中同一实体的数据元素之间的相对时序关系	$X=A/B$ 式中： A＝满足时序性要求的数据集中元素的个数； B＝被评价的数据集中元素的个数

5.7　可访问性

可访问性评价指标定义见表 6。

表 6　可访问性评价指标

指标编号	指标名称	指标描述	计算方法
0601	可访问	数据在需要时的可获取性	$X=A/B$ 式中： A＝满足可访问性要求的数据集中元素的个数； B＝被评价的数据集中元素的个数
0602	可用性	数据在设定有效生存周期内的可使用性	$X=A/B$ 式中： A＝满足可用性要求的数据集中元素的个数； B＝被评价的数据集中元素的个数

附　录　A
（资料性附录）
数据质量评价过程

图 A.1 描述了数据质量评价过程。

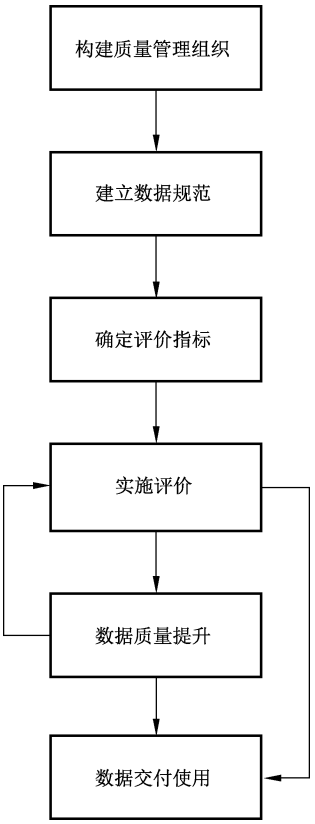


图 A.1　数据质量评价过程

参 考 文 献

[1] GB/T 5271.1—2000 信息技术 词汇 第1部分:基本术语
[2] GB/T 5271.17—2010 信息技术 词汇 第17部分:数据库