ICS 25.040.40
L 67

中华人民共和国国家标准

GB/T 39400—2020

工业数据质量　通用技术规范

Industrial data quality—General technical specification

2020-11-19 发布　　2021-06-01 实施

国家市场监督管理总局
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。

本标准由中国机械工业联合会提出。

本标准由全国自动化系统与集成标准化技术委员会(SAC/TC 159)归口。

本标准起草单位:中国标准化研究院、浙江大学、中机生产力促进中心、深圳鹏锐信息技术股份有限公司、深圳市华傲数据技术有限公司、北京三维天地科技股份有限公司。

本标准主要起草人:杨青海、王志强、顾复、洪岩、潘康华、刘守华、顾新建、岳高峰、肖承翔、张伟群、贾西贝、曹朝晖、徐凯程、尹书蕊。

引　言

随着信息化与工业化的深度融合，信息技术渗透到了工业企业产业链的各个环节，工业企业建立了很多计算机信息系统，积累了大量工业数据，工业数据已成为工业企业的重要资源。同时，数据质量贯穿于工业数据生命周期的产生、收集、存储、维护、传输、加工和利用等各个阶段，海量的工业数据存在数据残缺、数据不规范以及数据错误等数据质量问题。

本标准通过对工业数据质量持续改进的模型、质量的描述、识别、评价、控制和报告等的标准化，支撑工业数据的协同建设、互联互通、共享利用，提高工业数据的质量、可用性和利用效率。

本标准的实施有助于实现工业数据的规范化管理和质量保证。

工业数据质量　通用技术规范

1　范围

本标准规定了工业数据质量持续改进的模型，以及工业数据质量的描述、识别、评价、控制和报告的要求。

本标准适用于工业数据采集、传输、维护和使用过程中的质量管理。

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 19001—2016　质量管理体系　要求

3　术语和定义

下列术语和定义适用于本文件。

3.1

数据质量　data quality

数据的一组固有特性满足要求的程度。

注：固有特性一般指永久性的特性。

3.2

数据质量管理　data quality management

指导和控制某机构数据质量的协调活动。

3.3

质量评价过程　quality evaluation procedure

用于应用和报告质量评价方法及结果的操作。

3.4

质量测量　quality measurement

对质量定量元素、子元素的评估。

3.5

质量结果　quality result

数据质量测量得到的一个值或一组值，或将这些值同规定的一致性质量等级相比得到的评价结果。

3.6

质量范围　quality scope

报告质量信息的数据的覆盖范围或特征。

3.7

数据集　data set

可以标识的数据集合。

[GB/T 33674—2017，定义 3.1]

3.8

完全检查　complete inspection

质量范围内所有个体都进行的检查。

3.9

抽样检查　sampling inspection

从质量范围内的整体中抽取若干个体进行的检查。

3.10

主数据　master data

组织未来执行事务需要使用的，用于描述实体的独立的、基本的数据。

注 1：主数据通常包括描述客户、产品、雇员、材料、供应商、服务、股东、设施、设备以及规章制度的记录。

注 2：主数据的选择和确定，取决于组织的视角。

注 3：此处“实体”为一般含义，而非数据建模中使用的含义。

3.11

事务数据　transaction data

表征业务活动或活动方案实现的数据。

3.12

产品数据　product data

适合于人或计算机进行通信、解释或处理的，以形式化方法表达的有关产品的信息。

4　工业数据质量持续改进

4.1　质量特性

工业数据主要包括主数据、事务数据和产品数据。

工业数据质量特性包括完整性、一致性、准确性以及其他附加特性。

4.2　持续改进模型

工业数据质量管理应用戴明环（PDCA 循环）持续改进方法，PDCA 循环符合 GB/T 19001—2016 的规定，工业数据质量持续改进模型见图 1，包括策划、实施、检查和处置 4 个阶段，其中：

——策划（Plan）：明确质量目标和用户需求，规划数据质量描述要素，开展数据质量描述；

——实施（Do）：识别数据质量要素，新建数据质量要素，开展数据质量识别；

——检查（Check）：选择评价方法，确定评价流程和步骤，开展数据质量评价；

——处置（Act）：确定控制规则，选择控制方法，开展数据质量控制。

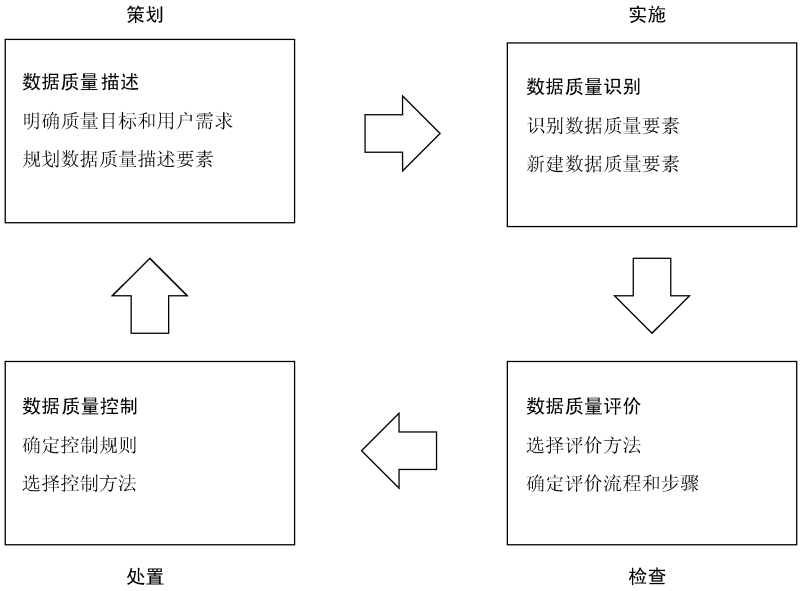

图 1　工业数据质量持续改进模型

5　工业数据质量描述

5.1　描述要素

源自数据集、用户需求的工业数据质量信息所反映的数据质量应满足用户的特定需求。质量目标表明数据质量应符合任务的特定目的。通过识别数据集、用户需求、质量目标中的质量元素来描述质量信息。质量描述可用于数据集系列、数据集或数据集内具有相同特征的部分数据。

一个数据集的质量用以下两个要素来描述：

——数据质量定量元素；

——数据质量非定量元素。

每个数据质量定量元素可细分为多个数据质量定量子元素。每个数据质量定量子元素用多个数据质量定量子元素描述项描述。通过数据质量定量元素、数据质量定量子元素及数据质量定量子元素描述项，描述数据集满足数据规范中预先设定标准的程度，并提供定量的质量信息。

数据质量非定量元素提供非定量的质量信息，可用于评价数据集在非预期的特定应用中的质量。

质量信息的可信性，记录在“数据质量报告”中。

数据质量描述框架见图 2。

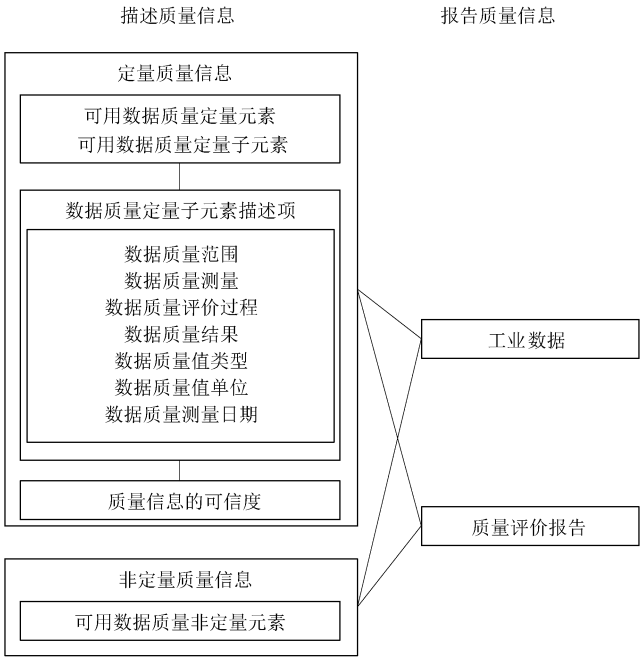

图 2　数据质量描述框架

5.2　定量元素

5.2.1　构成

数据质量定量元素用来描述数据集的定量质量信息，用来表达符合数据规范的程度。包括但不限于以下元素：

——完整性：特征、特征属性及特征关系存在或不存在；

——一致性：数据结构（包括概念结构、逻辑结构、物理结构）、属性及其关系符合逻辑规则的程度；

——准确性：包括正确性、精确性和时序性；

——附加数据质量定量元素：用户可根据需求设置，以便描述无法用以上定量元素描述的定量的数据质量信息。

5.2.2　子元素

数据质量定量子元素与数据质量定量元素相对应，用来描述数据集的定量质量信息。包括但不限于以下子元素：

a）　完整性的子元素：

——多余：数据集中有多余数据；

——缺少：数据集中缺少应有数据；

——交叉：数据集中存在交叉重复数据。

b）　一致性的子元素：

——概念一致性：符合概念模式规则；

——值域一致性：值在值域范围内；

——格式一致性：数据存储与数据集物理结构的一致性。

c）　准确性的子元素：

——正确性：数据反映和描述客观事物及其变化的准确程度；

——精确性：数值符合其实际值或规定值的程度；

——时序性：表达有序活动或序列活动相关数据时间顺序的正确性。

对任意数据质量定量元素，可新建附加数据质量定量子元素。

5.2.3 子元素描述项

对每个可用的数据质量定量子元素，应记录其质量信息。每个数据质量定量子元素的完全质量信息，用下列7个数据质量描述项来描述：

——数据质量范围；

——数据质量测量；

——数据质量评价过程；

——数据质量结果；

——数据质量值类型；

——数据质量值单位；

——数据质量测量日期。

5.3 非定量元素

数据质量非定量元素用来描述数据集的非定量的质量信息。包括但不限于以下元素：

——目的：描述数据集的创建原因和其预定的使用目的。

——用途：描述使用过该数据集的应用。数据生产者或其他数据使用者用“用途”来描述数据集的使用情况。

——数据志：描述数据集的历史，即数据集的整个生命周期信息。数据志包含两部分：描述数据集起源的源信息；描述数据集生命周期中的处理步骤和过程信息。数据溯源描述参见GB/T 34945—2017。

——附加数据质量非定量元素：描述以上数据质量非定量元素没有描述的非定量的质量信息。

6 工业数据质量识别

6.1 定量的数据质量信息

6.1.1 识别可用的数据质量定量元素

对可用于数据集的所有数据质量定量元素加以识别，判断这些元素是否适用于某一特定类型的数据集。

注：数据质量定量元素可用性由数据规范来决定。

6.1.2 新建附加数据质量定量元素

若本标准所列的数据质量定量元素未能充分描述数据质量的某一方面，则应命名并定义新的数据质量定量元素。附加数据质量定量元素的命名和定义，应作为数据集质量信息的一部分。

6.1.3 识别可用的数据质量定量子元素

对可用数据质量定量元素的所有数据质量定量子元素加以识别，判断这些元素的数据质量定量子元素是否适用于某一特定类型的数据集。每个可用数据质量定量元素至少包含一个可用数据质量定量子元素。

注：数据质量定量子元素可用性由数据规范来决定。

6.1.4 新建附加数据质量定量子元素

若本标准所列的数据质量定量子元素未能充分描述数据质量的某一方面，则应命名并定义新的数据质量定量子元素。附加数据质量定量子元素的命名和定义，应作为数据集质量信息的一部分。

6.1.5 数据质量定量子元素描述项使用

6.1.5.1 数据质量范围

对每个可用数据质量定量子元素，应识别至少一个数据质量范围。数据质量范围可以是数据集系列、数据集或数据集内具有相同特征的部分数据。若数据质量范围无法识别，则默认为该数据集。

注：数据质量范围的确定参照数据规范及数据质量非定量元素提供的非定量质量信息。

6.1.5.2 数据质量测量

每个数据质量范围有一个数据质量测量。数据质量测量应简要描述测量类型和测量边界。数据集的质量应由多个测量来衡量。

注：单一测量不能充分评价数据质量，也不能为数据集的所有应用提供单一测量。

6.1.5.3 数据质量评价过程

每个数据质量测量有一个数据质量评价过程。数据质量评价过程应描述数据质量范围内的数据质量测量方法，并包含该方法报告。

6.1.5.4 数据质量结果

每个数据质量测量有一个数据质量结果。数据质量结果应为以下两者之一：

——将数据质量测量应用到数据质量范围所限定的数据后得到的值或值的集合。

——将所得到的值或值的集合，用可接受的指定一致性质量等级，评价这些值或值的集合得到的结果。该数据质量结果为“通过”或“不通过”。

这两种类型的数据质量结果都应被提供。

6.1.5.5 数据质量值类型

每个数据质量结果有一个数据质量值类型。

注：“通过”或“不通过”的数据质量值类型为“布尔型”。

6.1.5.6 数据质量值单位

每个数据质量结果有一个数据质量值单位(若存在)。

6.1.5.7 数据质量测量日期

每个数据质量测量应有一个数据质量测量日期。

6.2 非定量的数据质量信息

6.2.1 识别可用的数据质量非定量元素

数据集目的应明确，用途应清晰，数据志应完整。

数据集的数据志应是可用的，或者报告数据志，或者报告缺少数据志的原因。

数据质量范围所限定的数据集内，当一部分数据的数据志与其他部分的数据志不同时，应提供其数

据志，作为非定量的数据质量信息完整记录的一部分。

6.2.2 新建附加数据质量非定量元素

若本标准所列数据质量非定量元素未能充分描述非定量数据质量的某一方面，则应命名并定义新的数据质量非定量元素。附加数据质量非定量元素的命名和定义，应作为数据集质量信息的一部分。

7 工业数据质量评价

7.1 评价方法

7.1.1 数据质量评价方法分类

数据质量评价方法分为：

——直接评价方法：通过比较数据与内外部参考信息来确定数据质量；

——间接评价方法：使用与数据相关的外部知识推断或估计数据质量。

7.1.2 直接评价方法

直接评价方法可分为：

完全检查方法：测试数据质量范围内的所有数据项；

抽样检查方法：测试数据质量范围内的部分数据项，抽样方法、抽样率及抽样过程应在数据质量报告中报告。

注：使用抽样检查方法时，特别是在使用小样本或非随机抽样时，分析数据质量结果的可信度。

7.1.3 间接评价方法

间接评价方法所依据的外部知识包括但不限于：数据质量非定量元素、数据集的其他质量报告。

注：仅当直接评价方法不可用时，才用间接评价方法。

7.2 评价流程和步骤

数据质量评价过程是产生和报告数据质量结果的一系列步骤。评价与报告数据质量结果的过程流见图 3，评价步骤见表 1。

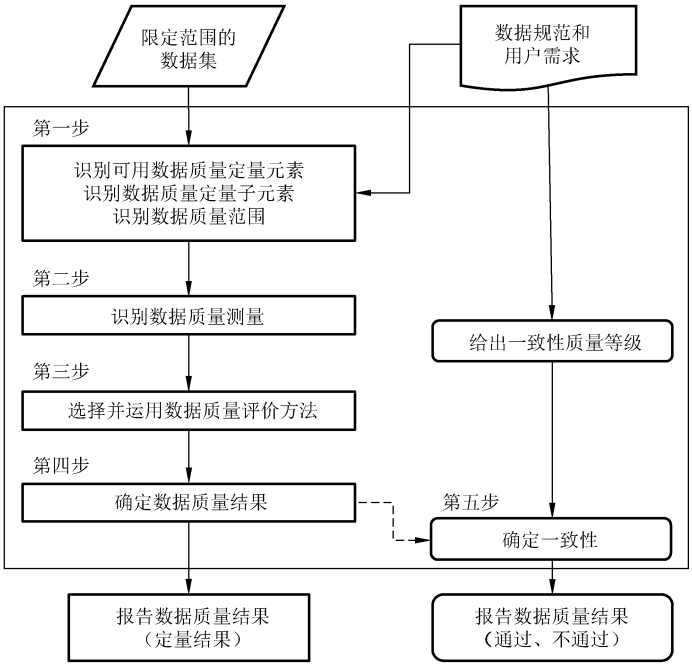

图 3　评价与报告数据质量结果的过程流

表 1　评价步骤

步骤	活动	描述
1	识别可用数据质量定量元素、数据质量定量子元素及数据质量范围	根据 5.2 识别数据质量定量元素、数据质量定量子元素及数据质量范围。若数据规范或用户需求有测试需要，则重复该步
2	识别数据质量测量	对每个测试，识别数据质量测量、数据质量值类型及数据质量值单位（若存在）
3	选择并运用数据质量评价方法	对每个被识别的数据质量测量，选择数据质量评价方法
4	确定数据质量结果	结果为：定量数据质量结果、数据质量值或数据质量值集合、数据质量值单位及数据质量测量日期
5	确定一致性	若数据规范或用户需求中已指定一致性质量等级，将其与数据质量结果相比可确定一致性。一致性数据质量结果（“通过”或“不通过”）是定量数据质量结果与一致性质量等级比较后的结果

8　工业数据质量控制

8.1　控制规则

8.1.1　数据质量描述测试套件

“数据质量描述测试套件”用来测试数据集质量描述的正确性。数据集质量描述应通过该测试套件的所有测试。

测试一：要素测试

a) 测试目的：证实质量要素都在质量描述中；

b) 测试方法：

——检查质量描述，证实数据质量定量元素、数据质量定量子元素及数据质量定量子元素描述项已被用来描述定量的质量信息；

——检查质量描述，证实数据质量非定量元素已被用来描述非定量的质量信息。

测试二：可用性测试

a) 测试目的：证实质量描述的可用性；

b) 测试方法：

——识别数据规范中与定量质量相关的语句，并用其来识别可用的数据质量定量元素及其可用的数据质量定量子元素；

——比较规范中识别的数据质量定量子元素与质量描述中所用的数据质量定量子元素，确保该数据集可用的所有数据质量定量子元素已被识别并用在质量描述中；

——检查可用的数据质量非定量元素，确保已被识别并用在质量描述中。

测试三：排斥性测试

a) 测试目的：证实质量描述中附加元素是排斥性的，且附加元素的信息已被充分提供；

b) 测试方法：

——检查所有附加数据质量定量元素，证实每个附加元素都描述了本标准中数据质量定量元素没有描述的定量质量信息；

——检查所有附加数据质量定量子元素，证实每个附加子元素都描述了本标准中数据质量定量子元素没有描述的定量质量信息；

——检查所有附加数据质量非定量元素，证实每个附加元素都描述了本标准中数据质量非定量元素没有描述的非定量质量信息。

测试四：定量子元素描述项正确性检查

a) 测试目的：证实数据质量定量子元素描述项使用正确；

b) 测试方法：比较本标准及每个可用数据质量定量子元素（包括附加数据质量定量子元素）所提供的质量信息，证实数据质量定量子元素描述项的使用符合本标准。

测试五：“数据质量报告”符合性检查

a) 测试目的：证实质量描述已用“数据质量报告”报告；

b) 测试方法：比较质量信息和“数据质量报告”，证实质量信息已用符合本标准要求的“数据质量报告”报告。

8.1.2 数据质量内容测试套件

8.1.2.1 测试目的：保证纳入“工业数据”的数据内容的质量。

8.1.2.2 测试方法：任何纳入“工业数据”的数据应符合给定的数据规范，并提供一致性数据质量报告，且在这些数据规范上的数据质量结果均为“合格”。一致性测试参见 GB/T 16656.31。

8.2 控制方法

数据质量控制总体上可分为三个步骤：

a) 生产者自查：生产者（数据集生产者）自查认为数据及其质量描述完全符合“数据质量描述测试套件”“数据质量内容测试套件”的所有要求，才能将其提交给第三方检查。

b) 第三方检查：第三方检查认为生产者提交的数据及其质量描述完全符合“数据质量描述测试套件”“数据质量内容测试套件”的所有要求，才能将其提交给评审组检查。否则，详细指出错误，将材料返回给生产者修改。

c) 评审组检查：评审组检查认为生产者提交的数据及其质量描述完全符合“数据质量描述测试套件”“数据质量内容测试套件”的所有要求，才能将其纳入“工业数据”。否则，详细指出错误，将材料返回生产者修改。

9 报告数据质量信息

9.1 概述

数据质量信息应按规范要求报告。

质量信息应以“数据质量报告”报告。

当多个数据质量结果被综合成单个数据质量结果来报告数据集质量时，综合数据质量结果应包含在“数据质量报告”中，其数据质量结果类型为“综合”。

9.2 数据质量报告

数据质量报告主要内容见表2。其中：

a) 编号：给表中每个条款编号。

b) 名称：报告条款名称。

c) 说明：定义或描述条款内容。

d) 约束/条件：描述报告该条款的必要条件，或需要该条款的条件。其含义如下：

——必备：应有该条款；

——条件可选：规定条件被满足时应有该条款；

——可选：该条款是可选的。

表2 数据质量报告主要内容

编号	名称	说明	约束/条件
1	质量报告	报告章节	必备
1.1	报告名称	报告的名称	必备
1.2	报告范围	该报告所评价数据集的范围	必备
2	数据质量测量	报告章节	必备
2.1	数学描述	数据质量测量的数学描述	必备
2.2	数据质量值	数据质量测量的结果值	必备
2.3	数据质量值单位	数据质量测量结果值的单位或值类型	必备
2.4	可信度	计算或估计的数据质量测量的可信度	必备
2.5	可信度单位	可信度的单位或值类型	必备
3	一致性的可信度	报告章节	必备
3.1	一致性结果可信度	一致性结果的可信度	必备
3.2	一致性结果可信度单位	一致性结果可信度的单位或值类型	必备
3.3	参考文档	一致性评价所参考的文档信息	可选
4	质量评价方法信息	报告章节	必备
4.1	方法类型	质量评价方法类型(直接评价方法、间接评价方法)	必备

表 2（续）

编号	名称	说明	约束/条件
4.2	检查策略	所用检查策略信息(完全检查方法、抽样检查方法)	必备
5	数据质量评价方法	报告章节	必备
5.1	假定	开发和应用该数据质量评价方法的隐含假定信息	必备
5.2	处理算法	为确定数据质量结果,处理数据的算法	必备
5.3	参数信息	数据质量评价方法所用参数信息	可选
5.3.1	参数定义	所用参数定义	必备
5.3.2	参数值	所用参数值	必备
5.3.3	参数单位	所用参数值的单位	必备
5.4	完全检查方法	完全检查方法的信息	条件可选
5.4.1	完全检查过程	完全检查过程的详细描述	必备
5.4.2	数据项描述	数据项的内容定义	必备
5.4.3	参考文档	完全检查所参考的文档	可选
5.5	抽样检查方法	抽样检查方法信息	条件可选
5.5.1	抽样检查方法类型	抽样检查方法的类型信息	必备
5.5.2	抽样过程	抽样过程详细描述	必备
5.5.3	抽样比率	样本占群体的比率	必备
6	综合质量	报告章节	条件可选
6.1	综合质量值	综合质量结果值	必备
6.2	综合质量值单位	综合质量结果值的单位或值类型	必备
6.3	综合方法	综合方法详细描述	必备
6.4	时间	综合时间	可选
6.5	综合质量报告	针对综合质量的报告	可选
7	非定量质量	报告章节	可选
7.1	目的	数据集预定目的信息	必备
7.2	用途	数据集使用情况信息	必备
7.3	数据志	数据集生命周期信息	必备
8	其他	报告章节	可选

参 考 文 献

[1] GB/T 16656.31—1997 工业自动化系统与集成 产品数据的表达与交换 第31部分:一致性测试方法论与框架:基本概念

[2] GB/T 19000—2016 质量管理体系 基础和术语

[3] GB/T 33674—2017 气象数据集核心元数据

[4] GB/T 34945—2017 信息技术 数据溯源描述模型

[5] GB/T 36344—2018 信息技术 数据质量评价指标