

T/QBDA

青岛市大数据发展促进会团体标准

T/QBDA 3051-2024

数据质量评估规范

Evaluation indicators for Data quality

(征求意见稿)

2024 - 0x - xx 发布

2024 - 0x - xx 实施

青岛市大数据发展促进会 发布

青岛市大数据发展促进会

目次

目次	I
前言	II
引言	III
数据质量评估规范	1
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 数据指标框架	2
4.1 评价指标框架	2
4.2 评价指标表头	2
4.3 指标详细说明	3
5 评估方法	11
5.1 检核方法	12
5.2 检核方式	12
5.3 分值设计	12
6 评估流程	13
6.1 流程图	13
6.2 评估准备	14
6.3 规则确定	14
6.4 评估实施	14
6.5 结果核验	14
6.6 报告编制	14
6.7 报告审核	14
参考文献	15

前言

本标准由青岛市大数据发展促进会提出并归口。

本部分起草单位：。

本部分起草人：

版权声明

根据国家有关法律法规规定，团体标准享有版权，未经许可禁止复制和销售。未经合法授权，严禁任何单位和个人对标准出版物及相关工作文件进行复制、销售、传播和翻译出版，严禁任何单位和个人将标准的任何部分通过电子信息网络或制作成标准数据库用于传播。标准化服务机构和中介机构的检验、鉴定、认证、咨询、评估和培训等活动必须使用正版标准。标准化研究机构馆藏的标准文本必须通过合法渠道购买。

引言

2021年3月发布的《中华人民共和国国民经济和社会发展第十四个五年规划和二〇三五年远景目标纲要》中提出，要充分发挥海量数据和丰富应用场景优势，促进数字技术与实体经济深度融合，赋能传统产业转型升级，催生新产业新业态、新模式，壮大经济发展新引擎。随着企业业务增长和规模扩大，以及伴随着信息技术和相关基础设施的不断完善，数据已经呈现了爆发式的增长，多数传统企业也开始走上了数字化转型的道路。数据已成为各类企业的重要生产要素，助力企业精益化决策与高效化生产、运营效率和产业效率，以及降低服务成本。

2022年，中共中央、国务院印发《关于构建数据基础制度更好发挥数据要素作用的意见》（以下简称“数据二十条”），对构建数据基础制度作了全面部署，明确提出推进数据资产合规化、标准化、增值化，有序培育数据资产评估等第三方专业服务机构，依法依规维护数据资源资产权益，探索数据资产入表新模式等要求。2023年，中共中央、国务院印发《数字中国建设整体布局规划》，进一步指出要加快建立数据产权制度，开展数据资产计价研究等。然而，不是所有的数据都能成为资产，数据的价值与数据质量密切相关。

青岛市大数据发展促进会经广泛调研，依据国家数据生产要素配置和数据产业的相关文件精神，特组织制订本评估标准。

本标准支撑企业对数据质量的准确度量与量化评估，有力提升数据的质量、可用性和利用效率，助力实现数据的规范化管理和质量保证。

青岛市大数据发展促进会

数据质量评估规范

1 范围

本文件规定了数据质量评价的指标和实施规范。
本文件适用于指导数据的质量评价工作。

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 25000.12-2017 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第 12 部分：数据质量模型

GB/T 25000.24-2017 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第 24 部分：数据质量测量

GB/T 35295-2017 信息技术 大数据 术语

GB/T 36344-2018 信息技术 数据质量评价指标

3 术语和定义

下列术语和定义适用于本标准。

3.1

数据 data

数据是指对客观事件进行记录并可以鉴别的符号，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合。根据《中华人民共和国数据安全法》的定义，数据是指任何以电子或者其他方式对信息的记录。

3.2

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， 2.3]

3.3

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。

[来源：GB/T 36344-2018]

3.4

数据生命周期 data life cycle

数据从数据的收集、存储、使用、加工、传输、提供、公开、删除等各种生存形态的演变过程。

4 数据指标框架

4.1 评价指标框架



图 1 指标框架

数据质量评价指标体系如图 1 所示，具体内容简要说明如下：

- 数据规范性：数据符合数据标准、数据模型、业务规则、元数据或者权威参考数据的程度。
- 数据完整性：按照数据规则要求，数据元素被赋予数据值的程度。
- 数据准确性：表示数据所描述的实体真实值的程度，如数据内容正确性、格式规范性等。
- 数据一致性：数据与其他特定上下文中使用的数据无矛盾的程度。
- 数据时效性：数据在时间变化中的正确程度。
- 数据可访问性：数据能被访问的程度。

4.2 评价指标表头

指标详情说明中表头信息详情说明如下：

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。

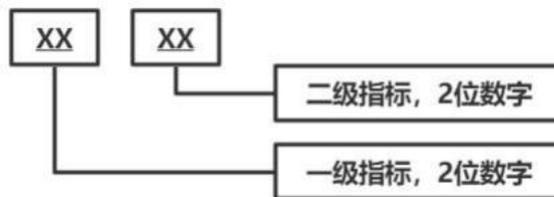


图 2 编码规则

1) 一级指标：由 2 位数字组成，详情见表 1。

表 1 一级指标编码及含义

编码	含义
01	数据规范性
02	数据完整性
03	数据准确性
04	数据一致性
05	数据时效性
06	数据可访问性

2) 二级指标：由 2 位数字组成的顺序码，范围 01~99

- b) 指标名称：评价指标的名称。
- c) 指标释义：评价指标的解释。
- d) 实现方式：评价指标如何度量的方式。
- e) 计算方法：评价指标的计算方法。
- f) 指标样例：对评价指标进一步说明的样例。
- g) 指标检测说明：评价指标检测进一步说明。

4.3 指标详细说明

4.3.1 数据规范性

数据的规范性，从三个方面来评价，分别是业务规范、技术规范和安全与隐私保护规范。

- a) 业务规范的要求：一是数据符合标准化要求的程度，标准化要求包括国际标准、国家标准、行业标准和地方标准等；二是数据符合权威参考要求的程度，权威参考主要指国家、行业或地方上权威机构发布的正式文件；三是数据符合组织业务规则要求的程度，比如某企业根据业务需要制定的数据规范要求。
- b) 技术规范包含数据模型和元数据，评价数据的组织形式、定义符合数据模型的度量和元数据的定义。
- c) 安全与隐私保护规范是指数据符合安全要求和隐私保护的程度，包括数据权限管理、数据脱敏处理等。

数据规范性评价指标详情见表 2。

表 2 数据规范性评价指标

指标编码	指标名称	指标释义	实现方式	计算方式	指标样例	指标检测说明
0101*	数据标准	数据符合数据标准的度量。	数据标准检测：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。 旧数据处理检测：数据规则中旧数据的销毁要具有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	例如： a、居民身份证号，在创建、命名、定义的时候，需要符合 GB11643-1999 《公民身份证号码》的要求； b、工商登记的企业数据，在企业注销的时候，需要符合工商信息化建设标准的要求。	需提供业务数据标准规范。 未制定相关数据标准的企业，此指标不适用，默认不具备此指标检测条件。
0102*	数据模型	数据符合数据模型的度量。	数据组模型检测：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	例如： a、公司基本信息表和法人信息通过法人身份证号关联，法人信息中必定存在身份证号。	需提供相关数据模型规范； 未制定相关数据模型规范的企业，此指标不适用，默认不具备此指标检测条件。
0103*	元数据	数据符合元数据定义的度量。	元数据文档检测：检查是否提供可解读的元数据文档，包括字段名称、描述、类型值域等内容的数据字典。	$X=A/B$ 式中： A=数据集中满足元数据要求的元素个数； B=数据集中数据元素总个数	例如： a、表和字段的命名需要遵循数据字典中的定义规范。	需提供相关元数据规范； 未制定相关元数据规范的企业，此指标不适用，默认不具备此指标检测条件。
0104*	业务规则	数据符合	业务规则文档	$X = A / B$	例如：	需提供相关业

		业务规则的度量。	检测：是否存在业务规则文档，及业务规则文档是否包含能够清晰描述业务交互的过程及结果数据记录。	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	a、满足业务要求的数据。	务规则文档；未制定相关业务规则的企业，此指标不适用，默认不具备此指标检测条件。
0105*	对权威参考数据（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。	标准码表检测：该数据集中元素如若涉及维度字段需包含相应的维表或解码说明； 字段值域检测：该数据集中元素的字段内容是否满足标准规范的值域要求。	$X = A / B$ 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	例如： a、中国 56 个民族，每个民族都有标准的代码和代码值，需要符合相关标准规范； b、行政区划代码要参考行政区划代码表。	需提供参考数据列表；未提供参考数据列表的企业，此指标不适用，默认不具备此指标检测条件。
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	数据分类分级检测：根据数据安全规范要求，是否对数据进行分类分级等安全处理； 数据脱敏检测：被列为敏感的数据集是否对相应数据元素进行脱敏处理。	$X=A/B$ 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	例如： a、法定代表人联系方式不应未被授权或脱敏就对外提供； b、法人的身份证件号码应满足国家个人信息保护法的要求。	取决于业务要求，法律法规的要求；未提供业务要求的企业，此指标不适用，默认不具备此指标检测条件。

注：指标编号中带有*的需提供相关文档

4.3.2 数据完整性

数据完整性，指是否存在构成信息的所有必要的的数据，不具备完整性的数据通常被看作已损坏或者数据丢失。数据完整性通常包括元素完整性和数据集完整性，其中元素完整性是指表中的一列没有缺失的程度，包括与业务信息相关的数据实体及实体属性没有缺失的程度；数据集完整性是指数据集中应出现的数据记录没有出现的程度。数据完整性评价指标详情见表 3。

表 3 数据完整性评价指标

指标编码	指标名称	指标释义	实现方式	计算方式	指标样例	指标检测说明
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	空值率检测：数据集中关键数据元素不为空的程度。	$X=A/B$ 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	例如： a、某城市一家空调品牌代理商的销售数据表中销售渠道、产品型号、安装方式、客户反馈数据元素缺失，不利于品牌厂商判断哪些渠道，不符合业务要求。	未有明确要求的，检测所有数据元素。
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	数据量检测：存储数据记录数与应存储数据记录数的百分比。	$X=A/B$ A=未被赋值的数据集中的数据量； B=预期被赋值的数据集中的数据量	例如： a、某城市一家空调品牌代理商的销售数据表，缺失 2022 年数据记录	需提供相关业务逻辑规则； 未提供业务逻辑规则的企业，此指标不适用，默认不具备此指标检测条件。

注：指标编号中带有*的需提供相关文档

4.3.3 数据准确性

准确性指标主要用于评价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，如数据内容正确性、数据格式准确性、数据重复率、数据唯一性、脏数据出现率等。数据准确性评价指标详情见表 4。

表 4 数据准确性评价指标

指标编码	指标名称	指标释义	实现方式	计算方式	指标样例	指标检测说明
0301*	数据内容准确性	数据内容符合预期的度量	先后关系内容准确：多个字段间具有先后关系； 检测数据内容	$X=A/B$ A=满足数据正确性要求的数据集中元素的个	例如： a、开始日期<=终止日期； b、某城市房产信息表中的装修情	需提供相关业务逻辑规则； 未提供业务逻辑规则的企业，此指标不适用，默认不具

			准确； 计算统计结果准确：一个值是由多个字段值计算出来，汇总前后内容准确。	数； B=被评价的数据集中元素的个数	况为“漏水”不符合预期值“精装修”、“毛坯”； c、总金额=基本金额+利息。	备此指标检测条件。
0302*	数据格式准确性	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	数据类型检测：数据元素内容类型是否符合预期； 数据长度检测：数据元素长度是否符合规范； 数据精度检测：对有精度要求的数据元素进行检测； 数据值域范围检测：数据是否处于该字段数据值域范围内。	$X=A/B$ 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	例如： a、某城市房产信息表中的房产面积值为100左右，不符合格式要求精确到小数后两位“98.00平”，竣工日期值为“25-9-13”，不符合日期格式要求 “yyyy-MM-DD”	需提供相关业务规则要求； 未提供业务规则要求的企业，此指标只检测日期格式、身份证号码18位、邮箱满足@.com格式、月份有12个月、年龄为数字、性别为男/女。
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	重复性检测：数据记录是否存在重复记录	$X=A/B$ 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	例如： a、某城市房产信息表中出现两条或多条房产编号为“F10000”的记录	出现重复的字段值均为重复数据。例如：100条数据中字段A的值50条数据为X，50条数据为Y，则重复数据为100，重复率为100%
0304*	数据唯一性	特定字段、记录、文件或数据集中单独字段唯一性的度量	字段唯一性检测：有唯一性要求的数据元素内容是否存在重复； 主键唯一性。	$X=A/B$ 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	例如： a、企业基本信息表中，明确了统一社会信用代码为主键，必须唯一，则不能出现统一社会信用代码相同的两条记	需要提供数据唯一性规则； 未提供关联规则的，此指标不适用，默认不具备此指标检测条件。

					录； b、企业资质基本信息中，同一类型的证书证书编号不能重复。	
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	关联表信息查询； 测试数据（带有测试字样或者test字样的数据）。	$X=A/B$ 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	例如： a、某城市房产已拆迁，此房产还有对应的业主信息，此业主数据为脏数据； b、测试时的遗留数据。	需要提供关联规则； 未提供关联规则的，此指标不适用，默认不具备此指标检测条件

注：指标编号中带有*的需提供相关文档

4.3.4 数据一致性

一致性指标主要用于评价表内或跨表数据的一致程度。包括相同数据一致性以及关联数据一致性。数据一致性评价指标详情见表 5。

表 5 数据一致性评价指标

指标编码	指标名称	指标释义	实现方式	计算方式	指标样例	指标检测说明
0401*	相同数据一致性	同一数据在不同位置存储或用户使用时，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	跨表数据内容一致性检测：数据集中关键元素在不同表中内容一致； 跨表数据格式一致性检测：数据集中关键元素在不同表中数据格式一致。	$X=A/B$ 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	例如： a、某集团企业采购合同数据：合同中的甲方信息数据一定是集团企业的标准基础数据。	需提供相关业务规则； 未提供业务逻辑规则的企业，此指标不适用，默认不具备此指标检测条件。
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	外关联约束类：引用其他业务对象属性时，所维护的属性值必须在其他业	$X=A/B$ A=满足一致性要求的数据集中元素的个数；	例如： a、外关联约束类：合同的签约客户必须	需提供相关业务规则； 未提供业务逻辑规则的企业，此指标不适用，默认不具备此指标检测条件。

			务对象中存在的约束； 跨表等值一致约束类：某一属性值与其他实体的一个或多个属性值的函数计算结果相等的约束； 跨表逻辑一致约束类：某一属性值满足其他实体的一个或多个属性值的函数关系的约束（大于或小于）。	B=被评价的数据集中元素的个数；	为客户主数据中定义的法人客户； b、跨表等值一致约束类：账单表的总金额与账单明细表的科目金额之和一致； c、跨表逻辑一致约束类：客户表中客户的入网日期早于客户订购产品表中的产品订购日期。	
--	--	--	--	------------------	---	--

注：指标编号中带有*的需提供相关文档

4.3.5 数据时效性

时效性指的是数据在时间变化中的正确程度。数据时效性评价指标详情见表 6。

表 6 数据时效性评价指标

指标编码	指标名称	指标释义	实现方式	计算方式	指标样例	指标检测说明
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	时间段记录数检测：时间段内的数据集数量是否符合业务要求； 时间段更新频率检测：在规定的周期内刷新。	$X=A/B$ 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	例如： a、医院系统中记录的某患者住院一周的血压监测情况，基于一周七天的日期范围，只有	需提供相关业务规则； 未提供业务逻辑规则的企业，此指标不适用，默认不具备此指标检测条件。

					六条记录，不符合基于时间段的数据正确性； b、产品价格必须每24小时刷新一次。	
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	检测真实业务发生的时间可能与该记录存在时间间隔是否符合业务需求	$X=A/B$ 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	例如： a、医院系统中记录的某患者住院一周的血压监测情况，在3月5号早上的血压测量时间为12:40，明显超过了早晨测量的时间06:00 - 10:00要求，不符合数据的及时性	需提供相关业务规则； 未提供业务逻辑规则的企业，此指标不适用，默认不具备此指标检测条件。
0503*	时序性	数据集中同一实体的数据元素直接的相对时序关系	抽取某一实体的相关记录，记录具有前后时序性	$X=A/B$ 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	例如： a、医院系统中记录的某患者住院一周的血压监测情况，抽取王某某的血压检测记录，未按照时间	需提供相关业务规则； 未提供业务逻辑规则的企业，此指标不适用，默认不具备此指标检测条件

					进行排序	
--	--	--	--	--	------	--

注：指标编号中带有*的需提供相关文档

4.3.6 数据可访问性

可访问性指的是数据能被访问的程度。数据可访问性评价指标详情见表 7。

表 7 数据可访问性评价指标

指标编码	指标名称	指标释义	实现方式	计算方式	指标样例	指标检测说明
0601*	数据可访问性	数据在需要时的可获取性	数据在需要时可以查询权限内允许的相关数据	$X=A/B$ 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	例如： a、数据审核员只能访问职责范围内的数据并进行审核	需提供相关业务规则； 未提供业务逻辑规则的企业，此指标不适用，默认不具备此指标检测条件
0602*	数据可用性	数据在设定有效生存周期内的可使用性	数据在被设定的有效生存周期内可以进行访问	$X=A/B$ 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；	例如： a、一家治疗某疾病的药品研发生产企业，每三年都会研发并向市场推出新产品，一些老药品也将会退市，那么企业ERP的退市药品相关数据在需要时可以随时调取分析	需提供相关业务规则； 未提供业务逻辑规则的企业，此指标不适用，默认不具备此指标检测条件

注：指标编号中带有*的需提供相关文档

5 评估方法

5.1 检核方法

数据质量评估按照评估指标对数据质量进行检核，检核方法包括：

- a) 系统检核：使用数据质量工具实现自动检核，如自定义SQL语句检核；
- b) 人工检核：根据评估指标，结合个人专业判断进行数据检核，如资料对比、经验判断等。

5.2 检核方式

5.2.1 抽样检核

按照抽样方案，对抽取的数据进行逐一检核数据质量。

5.2.2 抽样标准

数据表 数据量	参数	0-20	20-50	50-100	100-500	500-1000	1000 以上
0-10 万	表数量	100%	50%	25%	10%	5%	2%
	数据量	100%	100%	100%	100%	100%	100%
10 万-50 万	表数量	100%	50%	25%	10%	5%	2%
	数据量	65%	65%	65%	65%	65%	65%
50 万-100 万	表数量	100%	50%	25%	10%	5%	2%
	数据量	45%	45%	45%	45%	45%	45%
100 万以上	表数量	100%	50%	25%	10%	5%	2%
	数据量	15%	15%	15%	15%	15%	15%

5.3 分值设计

5.3.1 单一规则评价分计算

公式为：

$$G = 100 - \frac{Q}{S} \times 100$$

式中：

- Q —— 问题行数
- S —— 表总行数
- G —— 规则评估分

5.3.2 总分数计算

指标	规范性	完整性	准确性	一致性	时效性	可访问性
分数	EvaD	ComD	AccD	ConD	TimD	AvaD
权重	20%	15%	20%	15%	15%	15%

得分	EvaD × 20%	ComD × 15%	AccD × 20%	ConD × 15%	TimD × 15%	AvaD × 15%
----	------------	------------	------------	------------	------------	------------

注：权重值为建议值，可根据客户需求进行调整

公式为：

$$(EvaD \times 20\%) + (ComD \times 15\%) + (AccD \times 20\%) + (ConD \times 15\%) + (TimD \times 15\%) + (AvaD \times 15\%) = SumD$$

式中：

EvaD —— 规范性评估分

ComD —— 完整性评估分

AccD —— 准确性评估分

ConD —— 一致性评估分

TimD —— 时效性评估分

AvaD —— 可访问性评估分

SumD —— 总评估分

5.3.3 数据质量评分等级

依据本文件进行数据质量评价，对各项指标采取评分的方式予以打分评价，根据评分值评定数据资源质量水平，并以不同级别区分优质程度。按照从强到弱的要求分为不合格、合格、良好和优秀。数据质量评估级别及对应分值见表 8。

表 8 数据质量评估级别

评价级别	评价得分
优秀	90 分以上（含 90 分）
良好	75-90 分以上（含 75 分）
合格	60-75 分以上（含 60 分）
不合格	60 分以下（不含 60 分）

6 评估流程

6.1 流程图

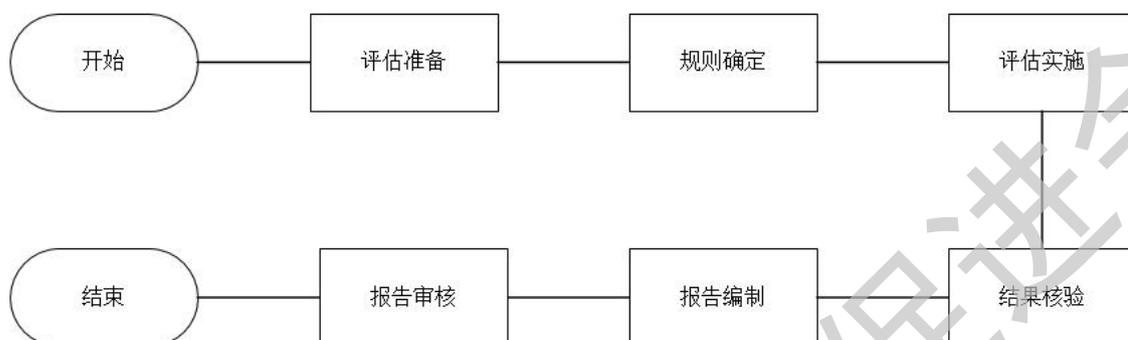


图 3 数据质量评估流程图

6.2 评估准备

评估前应了解具体业务对特定数据的需求，确定评估目的、对象及范围、评估方式，编制数据质量评估方案。

6.3 规则确定

根据评估指标、评估对象确定数据质量检核规则以及评估指标的权重。

6.4 评估实施

根据数据质量检核规则对数据进行数据质量检核。

6.5 结果核验

对检核结果进行核验，根据不同数据使用场景和业务需求对数据质量评估得分进行分级评价，确定数据质量好坏。

6.6 报告编制

根据评估结果编制评估分析报告，内容包括但不限于：评估对象及范围、评估指标、计分规则、评估检核方法、评估实施过程、质量问题。

6.7 报告审核

由报告审核员对报告进行核验，确保报告的准确性及合法性。并出具最终报告。

参 考 文 献

- [1] GB/T 25000.12-2017 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第 12 部分: 数据质量模型
- [2] GB/T 25000.24-2017 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第 24 部分: 数据质量测量
- [3] GB/T 35295-2017 信息技术 大数据 术语
- [4] GB/T 36344-2018 信息技术 数据质量评价指标