贵州省政府数据质量评价规范

Government Data Quality Assessment Specifications for Guizhou

ICS

**DB52**

贵 州 省 地 方 标 准

DB 52/T XXX—XXXX

2018 - XX - XX 发布 2018 - XX - XX 实施

贵州省市场监督管理局 发 布

目 次

[前言 1](#_Toc531252387)

[1 范围 1](#_Toc531252389)

[2 术语和定义 1](#_Toc531252390)

[2.1 数据质量维度 1](#_Toc531252391)

[3 数据质量评价原则 1](#_Toc531252392)

[3.1 科学性原则 1](#_Toc531252393)

[3.2 客观性原则 1](#_Toc531252394)

[3.3 系统性原则 1](#_Toc531252395)

[3.4 可操作性原则 1](#_Toc531252396)

[3.5 针对性原则 1](#_Toc531252397)

[3.6 引导性原则 2](#_Toc531252398)

[4 数据质量评价流程 2](#_Toc531252399)

[4.1 流程确定原则 2](#_Toc531252400)

[4.2 数据质量评价一般流程 2](#_Toc531252401)

[4.2.1 需求分析 2](#_Toc531252402)

[4.2.2 确定评价对象及范围 2](#_Toc531252403)

[4.2.3 选取评价维度及指标 2](#_Toc531252404)

[4.2.4 确定评价方法 2](#_Toc531252405)

[4.2.5 进行质量评价 2](#_Toc531252406)

[4.2.6 分析结果及报告 2](#_Toc531252407)

[5 数据质量评价维度 3](#_Toc531252408)

[6 数据质量评价指标 3](#_Toc531252409)

[6.1 主要数据质量指标 3](#_Toc531252410)

[6.2 评价指标选取的基本要求 4](#_Toc531252411)

[6.3 评价指标的筛选及权重 5](#_Toc531252412)

[6.4 评价指标的冲突处理原则 5](#_Toc531252413)

[7 数据质量评价方法 5](#_Toc531252414)

[7.1 定性方法 5](#_Toc531252415)

[7.1.1 第三方评测法 5](#_Toc531252416)

[7.1.2 用户反馈法 6](#_Toc531252417)

[7.1.3 专家评议法 6](#_Toc531252418)

[7.2 定量方法 6](#_Toc531252419)

[7.2.1 简单比率法 6](#_Toc531252420)

[7.2.2 最小/最大值法 6](#_Toc531252421)

[7.2.3 加权平均法 6](#_Toc531252422)

[7.3 综合方法 6](#_Toc531252423)

[7.3.1 主次分析法 6](#_Toc531252424)

[7.3.2 层次分析法 7](#_Toc531252425)

[7.3.3 模糊综合评价法 7](#_Toc531252426)

前 言

本标准按照GB/T 1.1—2009《标准化工作导则 第1部分：标准的结构和编写》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。

本标准由贵州省大数据发展管理局提出。

本标准由贵州省大数据标准化技术委员会归口。

本标准起草单位：云上贵州大数据产业发展有限公司、贵州中软云上数据技术服务有限公司。

本标准主要起草人：秦晓东、杨建国、王迪江、韦超、徐凯琳、李静、袁华、刘梦珠。

贵州省政府数据质量评价规范

1. 范围

本标准规定了贵州省政府数据质量评价的原则、一般流程、维度、指标、方法。

本标准适用于贵州全省范围内，用于指导政府部门对采集的结构化数据进行质量评估。

1. 术语和定义

下列术语和定义适用于本文件。

* 1. 数据质量维度

数据质量维度 Data Quality Dimensions

数据满足用户要求和使用目的的基本质量特性，是一个数据约束的类型。

1. 数据质量评价原则
   1. 科学性原则

评价应进行科学合理质量指标选择，以及采用科学合理的评价方法。评价必须有一定的理论作为基础，但又不能够脱离实际。

* 1. 客观性原则

评价应是符合实际、客观可信的。

* 1. 系统性原则

评价对象必须使用若干指标来衡量，指标间可能相互联系、相互制约，但每个指标又必须是独立的，不互相包容的，选取指标需考虑指标的层次性、系统性，避免指标间冲突。

* 1. 可操作性原则

指标的设计应避免过于繁琐，需考虑指标体系所涉及指标的量化及数据获取的难易程度和可靠性，评价体系应该是可行的、操作方便的。

* 1. 针对性原则

评价应能充分考虑各类政府数据资源所特有的类型特征并能将其揭示出来，要在指标的权重和分值上予以区分，以体现其针对性的导向作用。

* 1. 引导性原则

评价应以方便用户快捷而有效的选择和获取有价值的数据资源为导向。

1. 数据质量评价流程
   1. 流程确定原则

本规范提出数据质量评价的一般流程,具体的执行过程中根据数据对象或业务背景有所不同仍可根据具体情况适当增删。评价过程是一个迭代过程,各个过程的先后顺序仅表达阶段活跃的大致顺序,根据实际执行情况一些过程可能需要重复执行或省略。

* 1. 数据质量评价一般流程
     1. 需求分析

数据质量评价要以业务需求为中心进行。应了解具体业务对特定数据资源的需求特征，选择针对性的评价指标体系。

* + 1. 确定评价对象及范围

应根据需求分析结论选定当前用户为解决业务问题的数据作为评价对象，明确数据在属性、数量、时间等维度的具体界限。

* + 1. 选取评价维度及指标

在确定数据质量评价的对象和范围后, 应依据具体业务需求选择适当的数据质量维度和评价指标。要选取可测、可用的质量维度作为评价指标准则项。应注意指标之间避免冲突,同时也要注意新增评价指标的层次、权重问题,以及与其它同层次指标的冲突问题。

* + 1. 确定评价方法

数据质量评价在确定其维度和指标后,应该根据每个评价对象的特点,确定其评价方法,对于不同的评价对象需要选择不同的评价方法支持,应该根据质量对象的特点确定其评价方法。

* + 1. 进行质量评价

就是根据前面四步确定的质量对象、质量范围、评价方法实现质量评价的活动过程。

* + 1. 分析结果及报告

评测后要对评测结果进行分析：对评价目标与结果进行对比分析,确定是否达到评价指标；对评价的方案的有效性进行分析,确认是否合适等。

根据评价结果确定对象的质量评价,如需要,可根据评价结果鉴定质量级别。确定评价对象的质量级别是建立在相应的质量分级方案基础上的,该分级方案是根据相应的质量规范或用户的需求确定的,也是判断数据质量成熟度的重要依据。

质量评价结果和评价报告是所有具体业务数据质量评价项目及其评测结果的合集， 最后要将质量评价结果和数据质量评价过程汇总并报告。在完整的数据质量评价结果和报告中，应该包括全部上述内容。

1. 数据质量评价维度

数据质量是一个多维度的概念，涉及数据及其生产服务过程的多个方面。一般而言，对数据质量的认识通过将其分解为多个质量维度，并逐个识别实现。

数据资源具有用途个性化、多样化、不稳定等特点。数据质量要素受行业领域、数据类型和应用目的等因素的影响极大，不存在面向所有领域和资源类型的普适性数据质量指标体系，但针对一个具体行业背景下的特定数据类型，可以建立一组质量维度和指标体系。

按数据以及质量需求，本规范将数据质量评价要素划分为基本层、准则层和指标层三层。

基本层指标（一级指标）：按照数据的形式、内容和效用三个维度来认识，是数据质量评价的基本指标。

准则层指标（二级指标）：为数据质量评价活动中常见的质量指标，不保证其完备性，具体于一次评价活动中使用的质量指标仍需按照质量评价工作流程由评价方做出决定，。评价方应根据本次评价活动要求，可使用本规范以外的其他更加适用于数据资源的质量指标。

指标层指标（三级指标）：是总体评价目标的具体体现，是评价目标的具体分解。但每个指标都应从一定的角度或侧面反映评价目标，与评价目标紧密相关，同时指标总体对目标应具有足够的覆盖面，与评价目标保持高度的一致性。

1. 数据质量评价指标
   1. 主要数据质量指标

基于上述数据质量评价维度，本规范提出一组数据质量评价常用指标，供评价方在数据质量评价过程中选择使用。

数据对象范围的复杂，本规范仅将基本层质量指标作为要求性内容；而准则层指标为参考性内容，仅供评价方参考；未对指标层的具体执行方法进行约束，基本层和准则层主要指标如表1所示。具体的指标选取、权重确定和使用的评价方法由评价方在实施检查时依照质量评价流程确定。

表1 主要数据质量指标

| **质量指标** | | **说明** |
| --- | --- | --- |
| **基本层** | **准则层** |
| 形式质量 | 可获得性 | 可获得性就是数据资源获取的难易程度。反映数据资源查找及获得数据的便捷程度。 |
| 一致性 | 数据的一致性是指同一个数据在同一时刻只有一个值。反映数据资源描述要素、属性和它们间的相互关系符合逻辑规则的程度。 |
| 可理解性 | 可理解性是指阅读并理解数据资源编码的难易程度。反映数据资源描述的规范性以及数据内容的表达符合国内外相关规范、标准的程度。 |
| 完整性 | 数据资源完整性描述是对具体业务对象的信息完整程度。反映数据资源对实体描述的覆盖程度。 |
| 全面性 | 数据类目覆盖的全面程度。反映数据对业务范围的覆盖程度。 |
| 内容质量 | 准确性 | 准确性是数据所指内容对数据所指对象的反应、表现是否准确，以及数据形式对数据内容的表述、表达是否准确。反映数据对业务对象描述的准确程度。 |
| 正确性 | 正确性是指数据所指内容是否真实反映、表现出了数据所指对象的实际状况。反映数据对业务对象描述与实体对象的符合程度。 |
| 客观性 | 客观性是指数据采集和生产过程中是否受到主观因素影响以及被影响的程度。 |
| 可靠性 | 可靠性是指数据的“可信赖的”或“可信任的”程度，包括对数据采集、加工及数据应用等各环节的处理的可信赖程度。 |
| 安全性 | 确保数据安全所采取的安全措施和制定的安全机制。反映防止数据泄的防范程度。 |
| 效用质量 | 相关性 | 相关性是指数据描述的对象与业务实际对象之间的相关程度。反映指数据资源与用户需求匹配的程度。 |
| 共享/开放性 | 共享元目录中无条件共享/完全开放、有条件共享/依申请开放的总数占比率。反映数据资源的共享/开放程度。 |
| 效益性 | 效益性是指数据资源数据量是否足够实际业务使用，以及是否远多于用户需求的数量等。反映数据资源是否满足实际业务使用和使用是否对业务系统造成过量负荷（时间、精力、消耗）及数据的使用成本。 |
| 时效性 | 时效性主要考察数据的时间特性对业务应用的满足程度。不同类型的应用对数据的时间特性有不同的要求。反映数据资源更新的及时程度和对业务应用的支持程度。 |
| 活跃性 | 活跃性主要考查数据资源被调用和被访问的数据量。反映数据资源的贡献度、支撑度和数据关注热度。 |

* 1. 评价指标选取的基本要求

评价方决定一个数据质量指标时，应遵循以下原则：

* 指标选取要有系统性，以保证综合评价的全面性和可信度；
* 指标应意思明确，含义明确，不产生歧义；
* 选取的指标要有可测性，数据资料收集方便，计算简单，易于掌握。测量方法要长期保持有效；能被客观测量，而且能把数据质量在时间上做一个比较；
* 指标之间应尽可能避免明显的包含关系和相互冲突。对隐含的相关关系和相互冲突的指标，在模型中加以适当的消除和取舍；
* 指标的选择要保持同趋势化，以保证可比性；
* 指标设置要有重点，抓住主要因素。
  1. 评价指标的筛选及权重

评价指标是总体评价目标的具体体现，是评价目标的具体分解。不同数据资源的指标选择与权重侧重均应因对象不同而有所区别，但每个指标都应从一定的角度或侧面反映评价目标，与评价目标紧密相关，同时指标总体对目标应具有足够的覆盖面，与评价目标保持高度的一致性。

评价指标的选择应与数据的主要质量特征基本一致，最注重的应是数据的真实性、可达性和实用性方面的指标。

对于数据质量评价指标的分析主要通过专家调查分析法确定：

* 首先，通过统计分析得出适用频次较多的数据质量维度，结合数据资源的特点，选择若干个维度进行统一定义；
* 其次，从项目中对数据质量实际感知程度和理论认为重要程度两方面进行问卷调查；
* 结果处理：根据汇总调查结果，确定每项指标的权重；
* 根据实际情况截取权重较高的一定数量指标作为实际使用的评价指标，并根据数据资源的类型决定其测量方法。
  1. 评价指标的冲突处理原则

本规范要求的评价指标的分级与分类仅为实施评价活动而人为划分，因而有时会遇到指标间相互冲突的情况，对指标冲突的处理应遵循以下原则：

* 评价的有效性和评价的简便性矛盾之间应在满足有效性的基础上，尽可能简化；
* 评价的精确性和可信度矛盾之间，评价应尽可能精确，但目前有些指标不能做到很精确评价时，应保证指标的可信性，可请专家给出定性描述；
* 指标的系统性和指标的可获得性矛盾之间。因为指标体系要包括各个方面的许多因素，有些指标不易获得和不易测度，不能满足评价所需的全部数据。因此在建立指标体系时，对若干与评价关系不大的指标，虽然目前尚无法获得数据，仍以建议指标中提出，以保证评价指标体系的科学性和系统性。
* 在评价数据质量时存在准确性和可操作性之间的矛盾问题。从原则上讲，评价应保证一定的准确性，并具有一定的可操作性以使评价者依据评价指标能在相对统一的尺度上展开评价。但准确性与可操作性往往是矛盾的，并不能同时实现。因此可将评价指标体系考虑不同的要求，划分层次。

1. 数据质量评价方法
   1. 定性方法
      1. 第三方评测法

由第三方根据特定的信息需求，建立符合特定信息需求的数据质量评价指标体系，按照一定的评价程序或步骤，得出数据质量评价结论。

* + 1. 用户反馈法

由评价方向用户提供相关的评价指标体系和方法，用户根据其特定的信息需求从从中选择符合其需求的评价指标和方法来评价政府数据资源。

* + 1. 专家评议法

由若干专家组成的专家组来评价数据质量的过程。

* 1. 定量方法
     1. 简单比率法

简单比率法指期望的结果（E）占总值（T）的比率即E/T，反映企业（单位、其它机构等）数据质量某些方面的好坏程度。通常测量错误值（A）用“1–A/T”来表示。当结果等于或接近于1时，表明所反映的数据质量情况好，而当其等于或接近于0时，表明所反映的数据质量情况差。尽管一个期望结果的比率（E/T）与一个非期望结果（1–A/T）的比率给出的数据质量结果是相同的，但是E/T更有助于进行纵向比较，它可以反映数据质量在某些方面的改进。

* + 1. 最小/最大值法

适用于衡量数据质量中需要对多种指标进行加总的维度，评价的关键是要找出各类指标中的最小值或最大值。最小值法是一种保守的评估方法，它赋给维度一个不超过它最差数据质量指标的值。最大值是一种不保守的评估方法，一般适用于比较复杂的度量体系。

* + 1. 加权平均法

适用于在复杂的多指标评价时，评价者对于每个指标在总体评价中的重要程度很容易量化的场景。在综合评价操作前，为了确保评价值标准化，每个指标的权重必须被限定在0 和1 之间，并且它们的和等于1，即 ， 代表数据质量评价指标， 代表评价指标的权重，i=1,2,⋯，n。

* 1. 综合方法
     1. 主次分析法

是一种数学变换方法，把给定的一组相关变量通过线性变换转为另一组不相关的变量，这些新的变量按照方差依次递减的顺序排列。

主次分析法是根据评价指标中存在着一定相关性的特点，用较少的指标来代替原来较多的指标，并使这些较少的指标尽可能地反映原来指标的信息，从根本上解决指标间信息重叠问题，简化原指标体系的指标结构。

* + 1. 层次分析法

将与决策总是有关的元素分解成目标、准则、方案等层次，在此基础之上进行定性和定量分析的决策方法。层次分析法是将决策问题分解为不同的层次结构，通常按总目标、子目标层、准则层、方案层的顺序，然后利用求解判断矩阵特征向量的办法，求得每一层的各元素对上一层元素的优先权重，再用加权求和的方法递阶归并至各方案对总目标的最终权重，最终权重最大的方案即为最优方案。

层次分析法确定权重，是指采用1-9标度法，求出每层因素相对于评价目标的权重。首先将各层的指标的重要性比较判断，用1-9标度法构造标度判断矩阵；进行一致性检验，若检验通过，最大值特征值对应的特征向量即为权向量；若不通过，需重新构造标度判断矩阵。

* + 1. 模糊综合评价法

一种基于模糊数学的综合评标方法。该方法根据模糊数学的隶属度理论把定性评价转化为定量评价，即用模糊数学对受到多种因素制约的事物或对象做出一个总体的评价。它具有系统性强、结果清晰的特点,能较好地解决模糊、难以量化的问题,适合非确定性问题的解决。