ICS
J 98

T/CEC

中 国 电 力 企 业 联 合 会 标 准

T/CEC XXXXX—XXXX

电力数据质量评测规范

Specification for Quality Evaluation of Electric Power Data

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中国电力企业联合会　　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》编制。

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。

本文件由中国电力企业联合会提出。

本文件由电力行业 XXXX 技术委员会（XXXX）归口。

本文件起草单位：国家电网有限公司大数据中心、安徽继远软件有限公司、安徽继远检验检测有限公司、中国电力科学研究院有限公司、中国信息通信研究院、中国广核集团有限公司、中国华能集团有限公司、中国电力建设集团有限公司、华为技术有限公司、中国南方电网有限责任公司、数据易（北京）信息技术有限公司单位。

本文件主要起草人：。

本文件为首次发布。

本文件在执行过程中的意见或建议反馈到中国电力企业联合会标准化管理中心（北京市白广路二条一号，100761）。

电力数据质量评测规范

1 范围

本文件规定了电力数据质量评测的框架及评测过程。
本文件适用于电力数据生存周期各个阶段的数据质量评测。
本文件是对GB/T 36073数据质量域，数据质量评测维度的具体落地。
本文件是对GB/T 36344在电力领域的垂直细化。

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 GB/T 36344-2018信息技术　数据质量评价指标（所有部分）
GB/T 5271.1-2000 信息技术 词汇 第1部分：基本术语
GB/T 5271.17-2010 信息技术 词汇 第17部分：数据库

3 术语和定义

下列术语和定义适用于本文件。

3.1

数据 data

适用于通信、解释或处理的，信息再解释的形式化表示。
通过人工或自动手段处理数据。
[来源：GB/T 5271.1，有修改]

3.2

元数据 metadata

关于数据或数据元素的数据、数据描述，以及数据拥有权、存取路径、访问权和数据易变性的数据。
[GB/T 5271.17，有修改]

3.3

数据质量 data quality

在指定条件下使用时，数据特性满足明确的和隐含要求的程度。
[GB/T 36344，有修改]

3.4

原始数据 raw data

终端用户存储使用的未经处理或简化的数据。

原始数据有多种存在形式，包括文本数据，图像数据，音频数据或几种数据混合存在。

[GB/T 36344，有修改]

3.5

数据生存周期 data lifecycle

将原始数据转化为用于行动的知识的一组过程。

[GB/T 36344，有修改]

3.6

数据集 dataset

具有一定主题，标识并被计算机化处理的数据集合。

[GB/T 36344，有修改]

3.7

数据模型 data model

对分析图像和文本表述，识别组织为完成使命、功能、目标、目的和战略，以及管理和评价组织需要的数据。

从高到低的抽象层次表示数据时，区分概念模型、逻辑模型和物理模型。

数据模型使用周境边界正规描述，称为上下文模式。

数据模型标识实体、域或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，代表对业务有意义的人或行动事务集，以及描述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。

[GB/T 36344，有修改]

3.8

数据标准 data standard

数据命名、定义、结构和取值规范规则和基准。

[GB/T 36344，有修改]

3.9

业务规则 business rules

指数据符合业务流程、规则及要求的度量。

3.10

数据元 data element

也称数据元素，是用一组属性描述其定义、标识、表示和允许值的数据单元。

3.11

脏数据 dirty read

指源系统中的数据不在给定的范围内或对于实际业务毫无意义，或是数据格式非法，以及在源系统中存在不规范的编码和含糊的业务逻辑。

4 数据质量评测指标框架

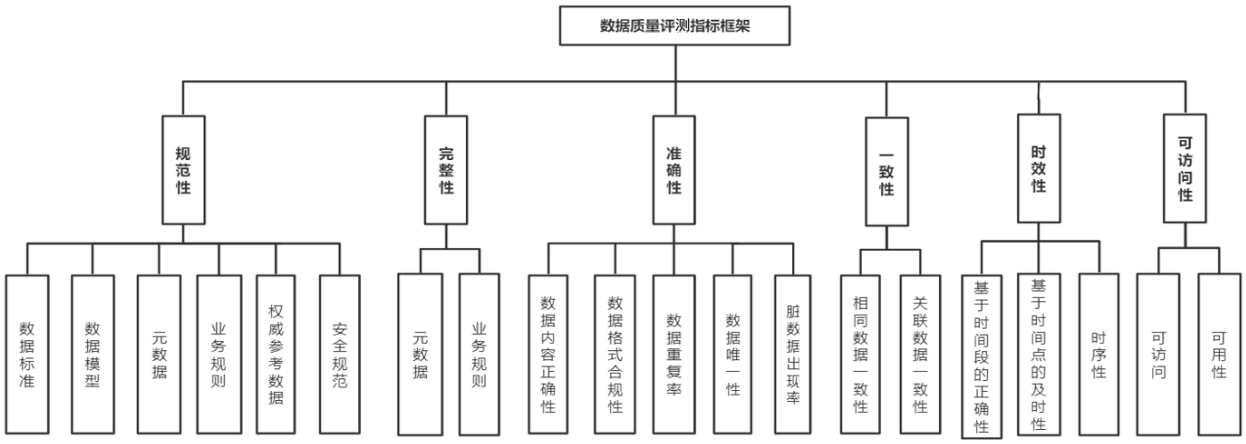

图 1 数据质量评测指标框架

数据质量评测指标框架见图 1。数据质量评测指标框架可包括下列内容：

a)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。

b)完整性：按数据规则要求，数据元素被赋予数值的程度。

c)准确性：数据准确表示描述真实实体或实际对象真实值的程度。

d)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。

e)时效性：数据在时间变化中的正确程度。

f)可访问性：数据被访问的程度。

5 数据质量评测指标

5.1 数据质量评测指标

数据质量评测指标可分为二级。一级指标应为数据质量评测指标框架中的六项指标，每项一级指标应包括数量不等的二级指标，每项二级指标应包括数量不等的指标子项。

5.2 数据质量评测二级指标

本标准详细规范了数据质量评测指标，对指标的计算进行了规范，详见附录 B。

5.3 数据质量评测二级指标及指标子项

本标准详细规范了数据质量评测指标二级指标及指标子项，对二级指标级子项进行了描述，详见附录 C。

6 数据质量评测过程

数据质量评测过程见图 2。

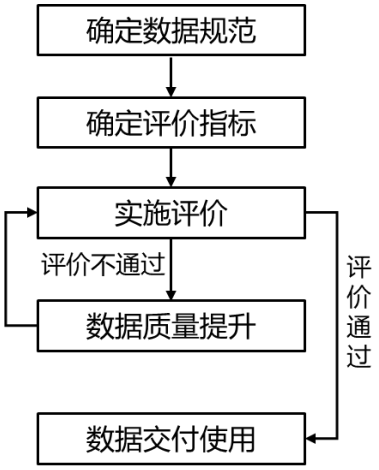

图 2 数据质量评测过程

6.1 确定数据规范

a）确定电力数据规范，经需求分析，确定评测对象及范围。

b）对业务数据质量评测应确定当前评估工作应用的数据集范围和边界，明确数据集在属性、数量、时间等维度界限。

c）评测对象可为数据项也可为数据集，但应为确定的静态集合。

6.2 确定评测指标

a）选取质量评测指标。应根据业务需求选择适当的评测指标，选取可测、可用的质量评测指标及指标子项。

b）根据实际需要，在不同数据类型和不同数据生产阶段，选取质量评测指标。

c）新增评测指标层次、权重问题，以及与其它同层次指标应避免冲突。

6.3 实施评测

a）确定质量测度及其评测方法。数据质量评测在确定维度和指标对象后，应根据每个评测对象特点，确定测度及实现方法。

b）不同评测对象测度和实现方法应根据质量对象特点确定。宜采用定性方法和定量方法。

c）质量评估，根据质量对象、质量范围、测量及其实现方法实现质量评测的活动过程。数据质量评测可采用下列方法：

d）自底而上：对数据集检查和评价，暴露数据分布分析、重复性分析、跨数据集依赖关系、孤岛数据记录和冗余分析等潜在数据异常和问题。自底而上方法，可评估出

异常、数据错误与业务背景无关。

e）自顶而下：业务用户参与记录业务流程和关键的数据依赖关系。在理解业务流程如何使用数据、哪些数据元素对业务应用至关重要的前提下评审。通过评审报告、记录和诊断的数据错误类型，评估与数据问题相关的业务影响。

6.4 数据质量提升

a）数据质量评测结果分析，清洗和校正数据质量缺陷。

b）确定和消除错误发生的根本原因；分离出不正确的数据项，采取符合预期的措施；可废除错误数据或纠正错误。

c）纠正错误方式可包括自动校正、人工指导校正、人工校正。

6.5 数据交付使用

符合质量要求的数据应交付数据应用使用。

7 电力综合数据质量评测

7.1 电力数据源

a）电力设备模型数据，智能电网调度控制系统中的设备模型数据质量影响着状态估计以及计算的结果。

b）电力稳态数据，在电网运行数据质量评测中，状态估计结果合格率可作为衡量指标，提取电力稳态数据指标时，应考虑影响潮流计算结果的因素，结合量测有效率、量测不平衡率、测量准确率来分析电力稳态数据。

c）电力故障特征数据，故障特征数据的质量如果不符合要求会导致故障误报的问题，给调度工作运行监控的决策管理带来了不良的影响，应根据数据实时率、正确率及完备率、可靠率指标分析电力故障特征数据。

7.2 电力综合数据评测方法

在综合评测中，多维度指标综合评分计算方法需要以各数据的考核权重为基础，考核权重的制定应基于主观权重与客观权重，其中主观权重指的是电气设备、量测装置的物理权重；客观权重指的是电气设备、量测装置各维度指标的变化属性。在实际的综合评测中，应对这两种权重因素进行综合考虑，使评价的结果能够反映电气的运行情况，因此需要计算出综合权重。

$$W_i = \frac{w_i^* w_i'}{\sum_{i=1}^{n} w_i^* w_i'}$$

式中，W_i为 i 维指标的综合权重；w_i^*为主观权重；w_i'为客观权重。

$$C = \begin{bmatrix} C_{11} & \cdots & C_{1i} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{ni} \end{bmatrix}$$

$$F = C \times W$$

式中，F为某类电网调度数据的综合评分数值；C为n个设备测量信号的i个维度的指标矩阵；W为i个考核指标的综合权重序列。综合评分数值换算为百分制，其中得分90-100为A级，75-90为B级，60-75为C级。

7.3 评测流程

首先，需要读取电力调度的基础数据，其中包括电气设备模型参数、电力稳态数据、电力故障特征数据。其次，需要读取各个维度指标的物理权重，还应进行指标计算，物理权重的设置可结合参与评测的设备、测量信号的重要性来确认。需要计算各个指标维度的变化熵，以熵值权重为基础来确认各指标维度的权重系数。最后，需要开展综合评分，将设备、量测、事件作为参考标准。评测的主要流程如下图所示。

图3 综合评测流程

附录 A

表A.1　数据质量指标体系权重分配表

综合评分	一级指标	二级指标	三级指标	四级指标（可计算）	是否纳入指标体系	评估方式	指标类型	计算方法	计算方式
总评（100%） 换算为百分制，得分90-100为A级，75-90为B级，60-75为C级。	规范性（10%）	数据标准(30%)	数据库标准化程度(30%)	是否有数据库规范要求(10%)	是	人工	自定义	1：有，0：没有	1：有，0：没有
				符合数据库命名规范的比例(10%)	是	系统	通用	符合数据库命名规范的数据库数量/数据库总数	不符合数据库命名规范的数据库数量/数据库总数
				符合数据库注释规范的比例(10%)	是	系统	通用	符合数据库注释规范的数据库数量/数据库总数	不符合数据库注释规范的数据库数量/数据库总数
				符合数据库索引命名规范的比例(10%)	是	系统	通用	符合数据库索引命名规范的数据库数量/数据库总数	不符合数据库索引命名规范的数据库数量/数据库总数
				符合数据库视图命名规范的比例(10%)	是	系统	通用	符合数据库视图命名规范的数据库数量/数据库总数	不符合数据库视图命名规范的数据库数量/数据库总数
				符合数据库序列命名规范的比例(10%)	是	系统	通用	符合数据库序列命名规范的数据库数量/数据库总数	不符合数据库序列命名规范的数据库数量/数据库总数
				符合数据库触发器命名规范的比例(10%)	是	系统	通用	符合数据库触发器命名规范的数据库数量/数据库总数	不符合数据库触发器命名规范的数据库数量/数据库总数

			符合数据库编码规范的比例(10%)	是	系统	通用	符合数据库编码规范的数据库数量/数据库总数	不符合数据库编码规范的数据库数量/数据库总数
			符合数据库密码命名规范的比例(10%)	是	人工	自定义	符合数据库密码命名规范的数据库数量/数据库总数	不符合数据库密码命名规范的数据库数量/数据库总数
			符合数据库备份规范的比例(10%)	是	人工	自定义	符合数据库备份规范的数据库数量/数据库总数	不符合数据库备份规范的数据库数量/数据库总数
		数据集标准化程度(30%)	是否有数据集规范要求(20%)	是	人工	自定义	1：有，0：没有	1：有，0：没有
			符合数据集命名规范的比例(20%)	是	系统	通用	符合数据集命名规范的的数据集数量/数据集总数	不符合数据集命名规范的的数据集数量/数据集总数
			符合数据集注释规范的比例(20%)	是	系统	通用	符合数据集注释规范的的数据集数量/数据集总数	不符合数据集注释规范的的数据集数量/数据集总数
			符合数据集容量规范的比例(20%)	是	系统	通用	符合数据集容量规范的的数据集数量/数据集总数	不符合数据集容量规范的的数据集数量/数据集总数
			符合数据集编码规范的比例(20%)	是	系统	通用	符合数据集编码规范的的数据集数量/数据集总数	不符合数据集编码规范的的数据集数量/数据集总数
		数据元标准化	是否有数据元规范要求(20%)	是	人工	自定义	1：有，0：没有	1：有，0：没有

		程度(40%)	符合数据元命名规范的比例(20%)	是	系统	通用	符合数据元命名规范的的数据元数量/数据元总数	不符合数据元命名规范的的数据元数量/数据元总数
			符合数据元注释规范的比例(20%)	是	系统	通用	符合数据元注释规范的的数据元数量/数据元总数	不符合数据元注释规范的的数据元数量/数据元总数
			符合数据元类型规范的比例(20%)	是	系统	通用	符合数据元类型规范的的数据元数量/数据元总数	不符合数据元类型规范的的数据元数量/数据元总数
			符合数据元容量规范的比例(20%)	是	系统	通用	符合数据元容量规范的的数据元数量/数据元总数	不符合数据元容量规范的的数据元数量/数据元总数
	数据模型(10%)	数据模型可定义性(30%)	有明确数据模型的定义的数据集比例(100%)	是	人工	自定义	有明确数据模型定义的数据集的数量/数据集总数	无明确数据模型定义的数据集的数量/数据集总数
		数据模型规范化程度(70%)	符合数据模型定义的数据集比例(100%)	是	人工	自定义	符合数据模型的数据集数量/数据集总数	不符合数据模型的数据集数量/数据集总数
	元数据(20%)	元数据规范性(30%)	是否有元数据规范要求(100%)	是	人工	自定义	1：有，0：没有	1：有，0：没有
		元数据文档(70%)	有元数据文档的数据集的比例(30%)	是	人工	自定义	有元数据文档的数据集数量/数据集总数	无元数据文档的数据集数量/数据集总数

			符合元数据定义的数据集比例(70%)	是	人工	自定义	根据现有情况制定元数据考核指标	不符合元数据定义的数据集数量/数据集总数
	业务规则(20%)	业务规则规范性(30%)	是否有业务规则规范要求(100%)	是	人工	自定义	1：有，0：没有	1：有，0：没有
		业务规则文档(70%)	有业务规则文档的数据集的比例(30%)	是	人工	自定义	有业务规则的数据集数量/数据集总数	无业务规则的数据集数量/数据集总数
			符合业务规则定义的数据集比例(70%)	是	人工	自定义	符合业务规则定义的数据集数量/数据集总数	不符合业务规则定义的数据集数量/数据集总数
	权威参考数据(10%)	参考数据可用性(100%)	已使用参考数据集的比例(100%)	是	系统	通用	已使用参考数据集数量/参考数据集总数	未使用参考数据集数量/参考数据集总数
	安全规范(10%)	安全规范性(30%)	是否有数据安全规范要求(100%)	是	人工	自定义	1：有，0：没有	1：有，0：没有
		数据权限规范(30%)	有数据权限控制的数据集的比例(100%)	是	人工	自定义	有数据权限控制的数据集的数量/数据集总数	无数据权限控制的数据集的数量/数据集总数
		数据脱敏(40%)	敏感字段标记的比例(50%)	是	系统	通用	已标记敏感字段的字段数据量/检测出的敏感字段的总数	未标记敏感字段的字段数据量/检测出的敏感字段的总数
			数据脱敏的比例(50%)	是	系统	自定义	敏感数据脱敏使用的次数/数据使用总次数	敏感数据未脱敏使用的次数/数据使用总次数

完整性（10%）	数据元素完整性（50%）	数据元缺失率（50%）	数据元缺失次数的倒数（100%）	是	人工	自定义	1（1+数据元缺失次数）	缺失数据元数量/数据元总数
		数据集缺失率（50%）	数据集缺失次数的倒数（100%）	是	人工	自定义	1/（1+数据集缺失次数）	缺失数据集数量/数据集总数
	数据记录完整性（50%）	空值率（50%）	空值率低于阈值的数据集的比例（100%）	是	系统	通用	空值率低于阈值的数据集的数量/数据集总数	空值记录数/数据记录总数
		数据缺失率（50%）	数据缺失率低于阈值的数据集的比例（100%）	是	系统	自定义	数据缺失率低于阈值的数据集的数据量/数据集总数	数据缺失数据集数量/数据集总数
准确性（30%）	数据内容正确性（30%）	数据内容规则正确性（50%）	符合标准数据元正则表达式的数据条目比例在阈值以上的数据集比例（100%）	是	系统	通用	符合标准数据元正则表达式的数据条目比例在阈值以上的数据集数量/存在标准数据元正则表达式的数据集总数	不符合正则表达式的数据记录数/数据记录总数
		数据内容语义正确性（50%）	具有语义可理解性的数据条目的比例在阈值以上的数据集比例（100%）	是	人工	自定义	具有语义可理解性的数据条目比例在阈值以上的数据集数量/数据集总数	语义难以理解的数据集数量/数据集总数

	数据格式合规性(30%)	数据类型合规性(25%)	符合数据类型要求的数据条目的比例在阈值以上的数据集比例(100%)	是	系统	通用	符合数据类型要求的数据条目的比例在阈值以上的数据集数量/数据集总数	不符合数据类型要求的数据元数量/数据元总数
		数值范围合规性(25%)	符合数值范围要求的数据条目的比例在阈值以上的数据集比例(100%)	是	系统	自定义	符合数据类型要求的数据条目的比例在阈值以上的数据集数量/数据集总数	不符合数值范围的记录数/数据记录总数
		数据长度合规性(25%)	符合数据长度要求的数据条目的比例在阈值以上的数据集比例(100%)	是	系统	自定义	符合数据类型要求的数据条目的比例在阈值以上的数据集数量/数据集总数	不符合数据长度的记录数/数据记录总数
		数据精度合规性(25%)	符合数据精度要求的数据条目的比例在阈值以上的数据集比例(100%)	是	系统	自定义	符合数据类型要求的数据条目的比例在阈值以上的数据集数量/数据集总数	不符合数据精度的记录数/数据记录总数
	数据重复率(15%)	数据重复率(100%)	数据重复率在阈值以下的数据集的比例(100%)	是	系统	通用	数据重复率在阈值以下的数据集的数据量/数据集总数	重复数据记录数/数据记录总数

	数据唯一性（15%）	数据唯一性（100%）	符合数据唯一性要求的数据集的比例（100%）	是	系统	通用	符合数据唯一性要求的数据集数量/有数据唯一性要求的数据集总数	不符合数据唯一性要求的数据记录数/数据记录总数
	脏数据出现率（10%）	脏数据出现率（100%）	脏数据出现率低于阈值的数据集的比例（100%）	是	系统	通用	脏数据出现率低于阈值的数据集数量/数据集总数	脏数据出现率低于阈值的数据集数量/数据集总数
一致性（10%）	与权威源的一致性（50%）	主从数据集一致性（100%）	从数据集与主数据集匹配的比例（100%）	是	系统	自定义	与主数据集匹配的数据集数量/存在主数据集总数	与主数据集不匹配的数据集数量/存在主数据集总数
	关联数据一致性（50%）	数据元组合一致性（100%）	符合标准数据元组合的数据集的比例（100%）	是	系统	自定义	符合标准数据元组合的数据集数量/存在标准数据元组合的数据集总数	不符合标准数据元组合的数据集数量/存在标准数据元组合的数据集总数
时效性（10%）	基于时间段的正确性（35%）	存量数据规模单调性（30%）	数据规模存量符合时序单调性的数据集比例（100%）	是	系统	自定义	T 周期内，数据规模存量具备时序单调性的数据集数量/数据规模存量具备时序单调性的数据集总数	T 周期内，数据规模存量不符合时序单调性的数据集数量/数据规模存量具备时序单调性的数据集总数

		存量数据规模时序稳定性(30%)	数据规模存量符合时序稳定性的数据集比例(100%)	是	系统	自定义	T 周期内，数据规模存量具备时序稳定性的数据集数量/数据规模存量具备时序稳定性的数据集总数	T 周期内，数据规模存量不符合时序稳定性的数据集数量/数据规模存量具备时序稳定性的数据集总数
		增量数据规模时序稳定性(40%)	数据规模增量符合时序稳定性的数据集比例(100%)	是	系统	自定义	T 周期内，数据规模增量具备时序稳定性的数据集数量/数据规模增量具备时序稳定性的数据集总数	T 周期内，数据规模增量不符合时序稳定性的数据集数量/数据规模增量具备时序稳定性的数据集总数
	基于时间点的及时性(35%)	增量数据API及时性(40%)	增量数据符合及时性要求的API 比例(100%)	是	系统	自定义	增量数据产生时间与入库/采集时间时延符合时延要求的 API 数量/有数据及时性要求的 API 总数	增量数据产生时间与入库/采集时间时延不符合时延要求的 API 数量/有数据及时性要求的 API 总数
		增量数据集及时性(30%)	增量数据符合及时性要求的数据集比例(100%)	是	系统	自定义	增量数据产生时间与入库/采集时间时延符合时延要求的数据集数量/有数据及时性要求的数据集总数	增量数据产生时间与入库/采集时间时延不符合时延要求的数据集数量/有数据及时性要求的数据集总数

		增量数据文件及时性(30%))	增量数据符合及时性要求的数据文件比例(100%)	是	系统	自定义	增量数据产生时间与入库/采集时间时延符合时延要求的数据文件数量/有数据及时性要求的数据文件总数	增量数据产生时间与入库/采集时间时延不符合时延要求的数据文件数量/有数据及时性要求的数据文件总数
	时序性(30%)	增量数据时序性(50%)	增量数据符合时序性要求的数据集比例(100%)	是	系统	自定义	增量数据符合时序性要求的数据集数量/增量量数据有时序性要求的数据集总数	增量数据不符合时序性要求的数据集数量/增量量数据有时序性要求的数据集总数
		存量数据时序性(50%)	存量数据符合时序性要求的数据集比例(100%)	是	系统	自定义	存量数据符合时序性要求的数据集数量/存量数据有时序性要求的数据集总数	存量数据不符合时序性要求的数据集数量/存量数据有时序性要求的数据集总数
可访问性（30%）	可访问性(70%)	数据库可访问性(10%)	可访问数据库比例(100%)	是	系统	通用	可访问数据库数量/数据库总数，每日更新	不可访问数据库数量/数据库总数，每日更新
		数据集可访问性(10%)	可访问数据集比例(100%)	是	系统	通用	可访问数据集数量/数据集总数，每日更新	不可访问数据集数量/数据集总数，每日更新
		数据元可访问性(10%)	可访问数据元比例(100%)	是	系统	通用	可访问数据元数量/数据元总数，每日更新	不可访问数据元数量/数据元总数，每日更新

			API可访问性(40%)	可访问API比例(100%)	是	系统	通用	可访问API数量/API总数，每日更新	不可访问API数量/API总数，每日更新
			数据文件可访问性(10%)	可访问数据文件比例(100%)	是	系统	通用	可访问数据文件数量/数据文件总数，每日更新	不可访问数据文件数量/数据文件总数，每日更新
			系统可访问性(10%)	可访问系统比例(100%)	是	系统	通用	可访问系统数量/系统总数	不可访问系统数量/系统总数
			存储设备可访问性(10%)	可访问设备比例(100%)	是	系统	通用	可访问设备数量/设备总数	不可访问设备数量/设备总数
		可用性(30%)	数据库可用性(20%)	数据库可用性评分(100%)	是	人工	自定义	通过调查问卷的形式，百分制	通过调查问卷的形式，百分制
			数据集可用性(20%)	数据集可用性评分(100%)	是	人工	自定义	通过调查问卷的形式，百分制	通过调查问卷的形式，百分制
			API可用性(20%)	API可用性评分(100%)	是	人工	自定义	通过调查问卷的形式，百分制	通过调查问卷的形式，百分制
			数据文件可用性(20%)	数据文件可用性评分(100%)	是	人工	自定义	通过调查问卷的形式，百分制	通过调查问卷的形式，百分制
			系统可用性(20%)	系统可用性评分(100%)	是	人工	自定义	通过调查问卷的形式，百分制	通过调查问卷的形式，百分制

附录B

表B.1 数据质量评测指标

一级指标	二级指标	指标描述	计算公式	计算公式描述	规则示例
规范性	数据模型	数据符合数据模型的度量。	X=A/B*100%	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	——
	元数据	数据符合元数据定义的度量。	X=A/B*100%	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	包含字段名称、描述、类型值域等的数据字典，为元数据文档
	业务规则	数据符合业务规则的度量。	X=A/B*100%	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	——
	权威参考数据	数据符合参考数据定义的度量。 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参考的数值集合或分类表。	X=A/B*100%	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	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
	安全规范	数据符合安全规范的度量。 安全规范是安全和隐私的规则，包括数据权限管理，数据脱敏处理等	X=A/B*100%	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	——
完整性	数据记录完整性	按业务规则要求，数据集中应被赋值的数据记录的赋值程度。	X=A/B*100%	A=被赋值得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	对表指定字段非空值检测
准确性	数据内容正确性	数据内容是否真实、准确反映 “真实世界”实体数据。	X=A/B*100%	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	订单金额+税额=发票金额
	数据格式合规性	数据类型、数值范围、数据长度、精度等数据格式是否真实、准确反映 “真实世界”实体数据。	X=A/B*100%	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	性别一栏不能出现男/女以外的内容。身份证不能出现标点符号；以及对字符编码的一些限制
	数据	特定字段、记录重复的度	X=A/B*100%	A=重复的数据集中元素	因为数据增

一级指标	二级指标	指标描述	计算公式	计算公式描述	规则示例
	重复率	量。		的个数 B=被评价的数据集中元素的个数	量历史数据合并异常造成的数据重复记录
	数据唯一性	特定字段、记录唯一性的度量。	X=A/B*100%	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	供应商编码对应供应商名称的唯一。 公司编码对公司名称的唯一 订单编号对合同编号的唯一
一致性	相同数据一致性	同一数据在不同位置存储数据的一致性； 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*100%	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	贴源层表与共享层表之间数据比对。
	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*100%	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	宽表与源表之间的数据比对。
时效性	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*100%	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	——
	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*100%	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	——
	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*100%	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	——
可访问性	可访问	数据在需要时的可获取性	X=A/B*100%	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	——
	可用	数据在设定有效生存周期	X=A/B*100%	A=满足可用性要求的数	——

一级指标	二级指标	指标描述	计算公式	计算公式描述	规则示例
	性	内的可使用性		据集中元素的个数; B=被评价的数据集中元素的个数	

附录 C

表 c.1 数据质量评测二级指标及指标子项

一级指标	二级指标	指标描述	指标子项	指标子项描述	规则示例
规范性	数据模型	数据符合数据模型的度量。	数据模型相对业务领域的覆盖率	存在数据模型的业务领域占所有业务领域的比率	
			数据符合数据模型的覆盖率	相关业务域中的数据集中的数据符合数据模型的比率	
			数据模型更新迭代	数据模型随业务实际变化及时更新	
			数据随数据模型版本更新的及时率	相关业务域中的数据集中的数据随数据模型的升级而及时更新	
	元数据	数据符合元数据定义的度量。	元数据相对业务领域的覆盖率	存在元数据的业务领域占所有业务领域的比率	
			数据符合元数据定义的覆盖率	相关业务域中的数据集中的数据符合元数据定义的比率	
			元数据丰富、更新迭代	元数据随业务实际变化及时增加、更新、删除	
			数据随元数据定义版本更新的及时率	相关业务域中的数据集中的数据随元数据的升级而及时更新	
	业务规则	数据符合业务规则的度量。	业务规则相对业务领域的覆盖率	梳理出清晰业务规则的业务领域占所有业务领域的比率	
			数据符合业务规则定义的覆盖率。	数据集是否能清晰的反应业务逻辑，字段和取值的具体意义是否明确。	

一级指标	二级指标	指标描述	指标子项	指标子项描述	规则示例
			业务规则丰富、更新迭代	业务规则随业务实际变化及时增加、更新、删除	
			数据随业务规则定义版本更新的及时率	相关业务域中的数据集中的数据随业务规则的升级而及时更新	
	权威参考数据	数据符合参考数据定义的度量。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。	相关数据赋值符合参考数据定义的比率	有参考数据的相关数据的赋值在参考的数值集合或分类表里	
			参考数据准确率	用来参考的数值集合或分类表符合业务实际	
			参考数据及时更新率	用来参考的数值集合或分类表随业务实际的变化更新	
	安全规范	数据符合安全规范的度量。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	识别出的相关国际安全规范是否全面	识别出和电力企业相关的数据国际安全规范，形成列表，并随时间更新列表。	
			国际安全规范合规率	数据符合识别出的国际安全规范的比率	
			识别出的相关国家安全规范是否全面	识别出和电力企业相关的数据国家安全规范，并形成列表，并随时间更新列表。	
			国家安全规范合规率	数据符合识别出的国家安全规范的比率	
			识别出的相关行业安全规范是否全面	识别出和电力企业相关的数据国行业安全规范，形成列表，并随时间更新列表。	
			行业安全规范合规率	数据符合识别出的行业安全规范的比率	
完整性	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。	数据集中应被赋值的数据记录有赋值的比率。	对于数据信息记录缺失的检测，可以通过对比源库上的表数据量和目的库上对应表的数据量来判断数据是否存在缺	数据集对具体业务对象的覆盖程度，一个数据集的特定属性

一级指标	二级指标	指标描述	指标子项	指标子项描述	规则示例
				失	都被赋予了数值
			数据集中应被赋值的数据记录的赋值符合业务规则的比率。	数据集中有赋值的数据记录，赋值符合业务规则要求。	
			数据集中应被赋值的数据元素（字段）有赋值的比率。	对于字段信息记录缺失的检测，选择需要完整性检查的字段,计算该字段中空值数据的占比，表的主键及非空字段空值率为 0%。空值率越小说明字段信息越完善，空值率越大说明字段信息缺失的越多。	
准确性	数据内容正确性	数据内容是否是真实、准确反映“真实世界”实体数据。	错误值占比	数据记录的信息存在错误的比率	用于描述一个值与它所描述的客观事物的真实值之间的接近程度。
			异常值占比	数据记录的信息存在异常的比率	
	数据格式合规性	数据类型、数值范围、数据长度、精度等数据格式是否真实、准确反映 “真实世界”实体数据。	数据类型合规	数据类型满足预期要求	
			数值范围合规	数值范围满足预期要求	
			数据长度合规	数据长度满足预期要求	
			数据精度合规	数据精度满足预期要求	
	数据重复率	特定字段、记录重复的度量。	记录重复率	记录重复占总记录数的比率	
			字段重复率	字段重复占总字段数的比率	
	数据唯一性	特定字段、记录唯一性的度量。	特定字段唯一性		
			特定记录唯一性		

一级指标	二级指标	指标描述	指标子项	指标子项描述	规则示例
一致性	相同数据一致性	同一数据在不同位置存储数据的一致性； 数据发生变化时，存储在不同位置的同一数据被同步修改。	数据静态一致性	同一数据在不同位置存储数据的一致性。	
			数据动态一致性	数据发生变化时，存储在不同位置的同一数据被同步修改。	
	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。	——	把待检测的表作为主表，首先用户确定一致性检测的主表字段，然后选择需要给定检测的从表和从表字段，设置好主表和从表之间的关联项,关联项可为多个字段，但是关联项应拥有匹配值的相似字段。匹配关联之后检查主表和从表相同或类似字段字段值是否一致。	
时效性	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	——	数据仅在一定时间段内具有价值的属性。数据从生成到录入数据库存在一定的时间间隔，该间隔较久，可导致分析得出的结论失去借鉴意义。	
	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	——	数据仅在某特定时间点前或后具有价值的属性。	
	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	——		
可访问性	可访问	数据在需要时的可获取性	易于采集	是否易于采集，采集过程是否简单直接	描述实际业务需要的数据获取的难易程度。包括

一级指标	二级指标	指标描述	指标子项	指标子项描述	规则示例
			合适的存储方式	数据存储结构是否合适，是否便于二次使用	采集、清理、转化等多个环节。
			易于处理	数据处理过程计算复杂度是否可接受	
	可用性	数据在设定有效生存周期内的可使用性	——		