



中华人民共和国国家标准

GB/T xxxx—xxxx

信用监管数据质量评价规范

Quality evaluation index of credit supervision data

(征求意见稿)

2027-XX-XX 发布

2027-XX-XX 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.....	II
1 范围.....	1
2 规范性引用文件.....	1
3 术语和定义.....	1
4 信用监管数据特征.....	3
4.1 数据范围.....	3
4.2 数据形态.....	4
5 数据质量评价应用场景.....	4
5.1 常规信用监管业务场景.....	4
5.2 含有人工智能的信用监管业务应用场景.....	4
6 评价原则.....	5
6.1 符合性.....	5
6.2 全面性.....	5
6.3 针对性.....	5
6.4 可操作性.....	5
6.5 动态性.....	5
6.6 安全性.....	5
7 评价指标体系.....	5
7.1 指标体系框架.....	5
7.2 指标适用性说明.....	6
7.3 指标编码.....	6
8 评价指标.....	7
8.1 规范性评价指标.....	7
8.2 完整性评价指标.....	8
8.3 准确性评价指标.....	9
8.4 一致性评价指标.....	10
8.5 时效性评价指标.....	11
8.6 可访问性评价指标.....	12
8.7 AI 可用性指标.....	12
9 评价方法.....	13
9.1 评价原则.....	13
9.2 评价方法.....	13
9.3 指标权重.....	14
参考文献.....	15

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》的规定起草。

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专利的责任。

本文件由全国社会信用标准化技术委员会（SAC/TC 470）提出并归口。

本文件起草单位：

本文件主要起草人：

信用监管数据质量评价规范

1 范围

本文件规定了经营主体信用监管数据的范围、适用数据形态、适用业务场景、评价要求、数据质量评价指标体系、评价方法。

本文件适用于市场监管部门、其他政府部门、信用服务机构、信用信息应用单位等对信用监管数据质量进行评价。

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 22117 信用 基本术语

GB/T 36344 信息技术 数据质量评价指标

3 术语和定义

GB/T 22117、GB/T 36344界定的以及下列术语和定义适用于本文件。

3.1

信用主体 **credit subject**

参与信用交易、管理、服务等一系列相关活动的组织和个人。

[来源：GB/T 22117—2018，2.4，有修改]

3.2

信用监管 **credit supervision**

基于信用主体的信用状况实施的差异化管理方式。

[来源：GB/T 22117—2018，9.2，有修改]

3.3

数据 **data**

任何以电子或其他方式对信息的记录，可以适用于通信、解释或处理。

[来源：GB/T 36344—2018，2.1，有修改]

3.4

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。

[来源：GB/T 36344—2018，2.2]

3.5

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—2018，2.6]

3.6

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—2018，2.3]

3.7

数据质量评价 data quality evaluation

按照数据质量评价指标体系，采用适当的方法对数据质量进行评估,并形成数据质量评价结果的过程。

3.8

数据标注 data labeling

给数据样本指定目标变量和赋值的过程。

[来源：GB/T 42755—2023，3.1]

3.9

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：实际应用中通常会区分语义数据模型和逻辑数据模型。

[来源：GB/T 36344—2018，2.7，有修改]

3.10

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。

[来源：GB/T 36344—2018，2.8]

3.11

人工智能 artificial intelligence;AI

人工智能系统相关机制和应用的研究和开发。

[来源：GB/T 41867—2022，3.1.2]

3.12

业务规则 business rules

在信用监管业务中，对数据的归集范围、更新周期、覆盖要求等作出的具体规定。

注：业务规则由相关法律法规、部门规章和业务规范确定，并随监管实践发展动态调整。

4 信用监管数据特征

4.1 数据范围

数据范围本文件所指经营主体为参与信用交易、管理、服务等一系列相关活动的企业、个体工商户和农民专业合作社等。

信用监管数据来源主要分为三类：市场监管部门在依法履职中产生的经营主体信息、经营主体通过公示系统填报和公示的信息、其他部门通过公示系统归集和公示的经营主体信息。

a) 市场监管部门在依法履职中产生对经营主体的监管信息，主要包括：

——经营主体登记注册备案信息（含企业基本信息、主要人员信息、吊销信息、注销信息、自然人出资信息、非自然人出资信息、投资人认缴信息、分支机构备案信息、变更备案信息、历史名称信息、分期实缴信息等）；

——行政处罚与行政许可信息（含行政处罚基本信息、行政处罚决定书信息、行政处罚变更信息、行政许可基本信息、行政许可变更信息）；

——经营异常名录信息（含企业异常名录信息、企业异常名录详细信息等）；

——严重违法失信企业信息（含严重违法失信企业名单、严重违法失信企业详细信息等）；

——信用修复信息（含严重违法失信名单（含自然人）详细信息、列入决定书信息、列入严重违法失信名单的具体情况信息、行政处罚信用修复信息等）；

——抽检抽查信息（含抽查检查信息）；

——股权出质登记信息（含股权出质登记信息、股权出质变更信息）；

——重点领域信息（含重点领域企业分类信息、信用承诺信息、告知承诺不实信息等）；

——食品抽检信息（含食品抽查检查信息）；

——公示公告信息（含公示公告信息、抽查检查公告详情信息、经营异常公告批量名单信息、严重违法公告批量名单信息、无证无照公告信息等）。

b) 经营主体通过国家企业信用信息公示系统主动填报和公示的信息，主要包括：

——年报信息（含企业年报基本信息、企业年报股东及出资信息、企业年报对外投资信息、企业年报网站或网店信息、企业年报股权变更信息、企业年报对外提供保证担保信息、企业年报修改信息等）；

——企业公示即时信息（含出资人信息、出资人认缴明细信息、出资人实缴明细信息、股东及出资修改信息、股权变更信息、许可信息、行政处罚信息、许可变更信息、知识产权出质登记信息、知识产权出质变更信息、行政处罚变更信息）；

——执行标准自我声明信息（含执行标准自我声明信息、执行标准自我声明修改信息等）。

c) 其他部门通过公示系统归集和公示的经营主体信息，主要包括：

——双随机信息（含抽查计划及任务信息、检查工作信息、检查工作信息名录信息、抽查结果信息等）；

——其他部门行政处罚与行政许可信息（行政处罚基本信息、行政许可基本信息）；

——其他部门公示信息（含抽查检查信息、其他部门列入严重违法失信企业名单、其他部门列入严重违法失信企业名单（黑名单）详细信息等）；

——股权冻结信息（含股权冻结被执行人信息、股权冻结信息、股权变更信息）；

——跨部门联合监管信息、司法执行“扫码入企”信息、联合奖惩信息等。

注：1. 具体数据项可参考GB/T 22120《企业信用数据项》、GB/T 40478《企业信用监管数据规范》等相关国家标准。

2. 信用监管数据的具体内容、范围和要求由相关法律法规、部门规章和业务规范确定，并随监管实践发展动态调整。

4.2 数据形态

本标准规定的信用质量评价指标适用于信用监管领域的各类数据，包括结构化数据、非结构化数据和半结构化数据。

a) 结构化数据：包括关系型数据库中的表格数据、电子表格数据等，如经营主体基本信息表、行政处罚记录表、年报信息表、抽查检查结果表等；

b) 非结构化数据：包括文本文件、图像、音频、视频等，如行政处罚决定书扫描件、现场检查照片、执法记录视频、公示公告原文等；

c) 半结构化数据：包括XML、JSON、HTML等格式的数据，如通过API接口交换的信用信息、网页公示信息、双随机抽查计划数据等。

5 数据质量评价应用场景

5.1 常规信用监管业务场景

信用监管数据质量评价应覆盖信用监管业务的全过程和各环节，具体包括以下应用场景：

a) 信用信息归集、共享和公示：信用数据的采集、整合、归集；跨部门、跨地区信用数据共享交换；信用数据的公示和公开；

b) 信用评价和信用分级分类：公共信用综合评价、市场化信用评价、信用分级分类监管、信用风险预警等；

c) 信用承诺和信用核查：信用承诺信息归集、信用承诺履行核查、信用核查在行政审批和市场监管中的应用等；

d) 守信激励和失信惩戒：守信激励措施实施、失信惩戒信息应用、联合奖惩机制运行等；

e) 信用修复：信用修复申请处理、信用修复过程记录、信用修复结果归档等；

f) 信用监管数据质量监测和改进：信用数据质量日常监测、数据质量问题识别整改、数据质量持续改进等；

g) 其他信用监管应用场景：信用监管决策支持、信用监管效能评估等。

注：业务场景随监管实践发展动态调整。

5.2 含有人工智能的信用监管应用场景

a) 信用风险智能预警和分级分类：信用风险预测模型训练、信用风险等级自动分类、高风险主体智能识别和预警等；

b) 信用画像智能构建：基于多源数据融合的市场主体全景画像生成、信用特征智能提取、信用标签自动标注等；

c) 信用数据智能治理：信用数据自动清洗和标准化、信用数据质量智能检测和修复、信用数据元信息智能补充等；

d) 智能评标和招投标监管：基于人工智能的标书智能评阅、围串标行为智能识别、招投标信用风险智能预警等；

e) 信用监管智能决策支持：基于人工智能的监管策略优化、检查计划智能生成、监管资源配置智能推荐等；

f) 信用数据质量智能评估：基于人工智能的数据质量指标自动计算、数据质量问题智能诊断、数据质量改进建议智能生成等；

g) 其他人工智能应用场景：人工智能模型训练数据质量评价、人工智能应用效果评估等。

注：应用场景随监管实践发展动态调整。

6 评价原则

6.1 符合性

评价指标框架符合GB/T 36344中关于数据质量评价指标的基本要求，并结合信用监管特点进行细化和扩展。

6.2 全面性

评价应覆盖数据的全生命周期，包括数据采集、存储、处理、传输、应用等各个环节。

6.3 针对性

应根据数据的具体形态、业务场景和应用需求，选择适用的评价指标。

6.4 可操作性

评价方法应简便易行，评价结果应可量化、可比较。

6.5 动态性

应根据监管实践发展和业务需求变化，适时调整评价指标和权重。

6.6 安全性

应符合国家网络安全、数据安全和个人信息保护等相关法律法规要求。

7 评价指标体系

7.1 指标体系框架

信用监管数据质量评价指标体系主要包括规范性、完整性、准确性、一致性、时效性、可访问性以及AI可用性等七个评价维度。

规范性、完整性、准确性、一致性、时效性、可访问性六个基础维度遵循GB/T 36344的要求，是常规性指标；AI可用性是根据信用监管智能化发展趋势新增的扩展维度，用于评价数据对人工智能模型开发和训练的支撑能力。

信用监管数据质量评价指标体系见图1。

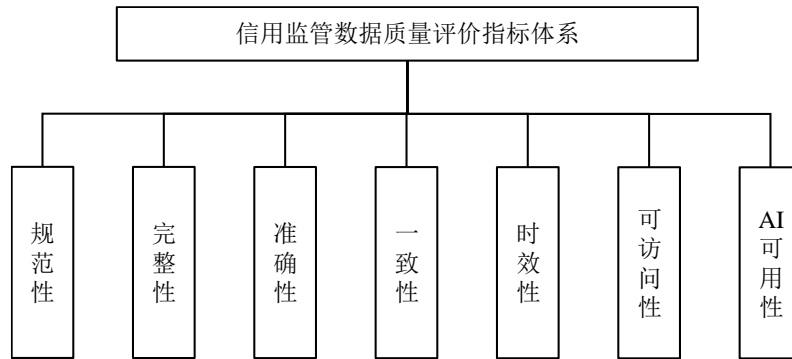


图1 信用监管数据质量评价指标体系

- a) 规范性：数据符合信用监管业务规则、技术标准和法律法规要求；
- b) 完整性：数据按照信用监管业务规则要求，包括数据元素完整性和数据记录完整性的程度；
- c) 准确性：数据所描述的实体真实值的程度，保证数据符合业务规定的合理取值和逻辑关系；
- d) 一致性：数据与其他特定上下文中，其内容、格式、关系等保持协调一致，无矛盾和冲突的程度；
- e) 时效性：数据应符合信息归集公示的时限和更新频率要求，在时间变化中的正确程度；
- f) 可访问性：数据能被授权用户或系统有效发现、获取和读取的程度；
- g) AI可用性：数据集在规模、分布、标注、时效等方面，符合用于训练信用监管领域人工智能模型特定要求的程度。

7.2 指标适用性说明

a) 常规性指标适用范围：常规性指标包括：规范性、完整性、准确性、一致性、时效性、可访问性六个基础维度，用于评价信用监管数据的基本质量特征。适用于所有信用监管数据质量评价场景，为必选指标。

b) AI可用性指标适用范围：AI可用性指标为人工智能可用性指标。用于评价数据对人工智能模型开发和训练的支撑能力，是结合信用监管智能化发展趋势新增的扩展维度。仅适用于具备人工智能模型开发、训练或应用需求的场景。执行部门可根据实际需要选择是否纳入评价范围。

7.3 指标编码

指标代码是评价指标的唯一性代码，采用层次编码方法，编码位数一般为4位，按照一级指标、二级指标的从属关系顺序编码。如需扩展，可延伸到三级指标，相应的编码位数调整为6位。每一级指标代码分别用2位阿拉伯数字表示。编码规则如图2所示。

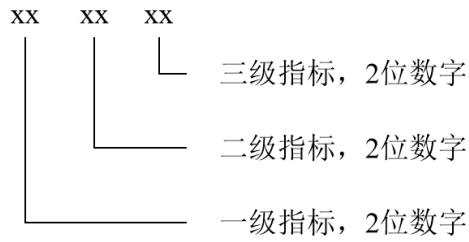


图2 指标编码规则

a) 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、07代表AI可用性。

b) 二级指标：由2位数字组成的顺序码，范围为01~99。

c) 三级指标（如适用）：由2位数字组成的顺序码，范围为01~99。

8 评价指标

8.1 规范性评价指标

信用监管数据规范性是指数据在结构、内容、管理及安全等方面，符合既定的信用监管业务规则、技术标准、管理规范及法律法规要求的程度。

规范性评价指标及计算方法见表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	数据命名规范性	数据项（如数据库表、字段、代码等）的命名和唯一标识符合相关业务标准、技术标准或管理规范的程度。	$X=A/B$ 式中： X——数据命名规范性指标得分； A——命名与标识符合规范要求的数据项数量； B——被评价的数据项总数。
0102	数据格式规范性	数据值的类型、长度、精度、模式等符合既定格式规范的程度。	$X=A/B$ 式中： X——数据格式规范性得分； A——格式符合规范要求的数据记录数或数据项数； B——被评价数据记录总数或数据项总数。
0103	数据值域规范性	数据项的取值在其定义的允许值集合内，且使用的代码符合权威代码标准的程度。	$X=A/B$ 式中： X——数据值域规范性得分； A——取值在标准值域内或使用标准代码的数据项数量； B——应使用标准值域或代码的数据项总数。
0104	数据模型规范性	数据的结构、关系及约束符合信用监管领域概念模型、逻辑数据模型等规范要求的程度。	$X=A/B$ 式中： X——数据模型规范性得分； A——符合数据模型规范要求的数据实体、属性或关系数量；

指标编号	指标名称	指标描述	计算方法
			B——被评价的数据实体、属性或关系总数。
0105	元数据规范性	元数据的完整性、准确性、规范性符合相关标准要求的程度。元数据包括数据名称、数据定义、数据类型、数据格式、数据来源、数据更新频率、数据责任部门、数据安全等级等描述信息。	$X=(C_1 \times W_1 + C_2 \times W_2)$ 式中： X——元数据规范性综合得分； C ₁ ——元数据完整率（已定义的必备元数据项数/应定义的必备元数据项总数）； C ₂ ——元数据准确率（内容符合规范的元数据项数/被评价的元数据项总数）； W ₁ , W ₂ ——分别为完整性和准确性的权重，建议值 W ₁ =0.4, W ₂ =0.6, 总和为 1。 注：可根据实际情况调整评价维度和权重。
0106	业务规则合规性	数据内容满足信用监管特定业务规则、逻辑约束和政策要求的程度。	$X=A/B$ 式中： X——业务规则合规性得分； A——符合所有适用业务规则的数据记录数； B——被评价的数据记录总数。
0107	权威参考符合性	数据内容与法定、权威数据源或经确认的参考数据保持一致的程度。例如，企业登记信息与国家企业信用信息公示系统权威库比对。	$X=A/B$ 式中： X——权威参考符合性得分； A——与权威参考源一致或经其验证通过的数据项数量； B——应进行权威参考比对的数据项总数。
0108	数据安全合规性	数据的采集、存储、传输、访问、使用、销毁等全生命周期处理活动符合国家数据安全法律法规、网络安全标准的程度。	采用符合性检查法。 $X=(S/T)$ 式中： X——数据安全合规性得分； S——已实施并符合要求的安全措施项数； T——应实施的安全控制措施项总数。
0109	隐私保护合规性	数据中涉及个人隐私信息的收集、存储、使用、加工、传输、提供、公开、删除等全生命周期处理活动应符合《个人信息保护法》《数据安全法》《敏感个人信息处理安全管理要求》等法律法规和标准规范的程度。	采用符合性检查法。 $X=(S/T)$ 式中： X——隐私保护合规性得分； S——已实施并符合要求的隐私保护措施项数； T——应实施的隐私保护控制措施项总数。

8.2 完整性评价指标

信用监管数据完整性是指数据在归集和提供过程中，按照业务规则要求所必需的数据项、数据记录及数据集合完备、无缺失的程度，主要衡量信用信息归集要求的达成情况。

完整性评价指标及计算方法见表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	数据项完备性	数据集中单条数据记录内按照业务规则要求必须填写的核心数据项实际被赋值且非空的比例。	$X=A/B$ 式中： X——数据项完备性得分； A——被评价数据集中，所有记录的核心必填数据项非空的总数； B——被评价数据集中，所有记录的核心必填数据项应填总数。
0202	记录完备性	在特定时间周期或业务场景下，实际归集到的有效数据记录数量与理论上应存在的目标数据记录总数的比例。	$X=A/B$ 式中： X——记录完备性得分； A——实际归集到的有效数据记录数； B——根据业务规则，在评价范围内应存在的目标数据记录总数。
0203	数据集完备性	在某一信用监管主题或业务链条下，实际可用的完整数据集数量与按规定应提供的完整数据集数量的比例。	$X=A/B$ 式中： X——数据集完备性得分； A——实际提供的、符合质量要求的完整数据集数量； B——按规定应提供的完整数据集总数。
0204	业务链条完整性	针对特定经营主体，其全生命周期或特定业务链条中各环节所产生的关键数据记录，包括本周期内时间覆盖全面性和部门事项覆盖全面性要求。被连续、完整归集的比例。	$X=A/B$ 式中： X——监管链条完整性得分； A——被评价主体集合中，拥有完整监管链条记录的主体数量； B——被评价的主体总数。
0205	记录唯一性	数据集中，不存在因归集或处理错误导致的完全重复或关键信息重复的数据记录的比例。	$X=1-D/T$ 式中： X——记录唯一性指标得分； D——被判定为无效重复的数据记录数； T——被评价的数据记录总数。

8.3 准确性评价指标

信用监管数据准确性是指数据内容与其所描述的信用主体真实属性、状态或发生事件的实际状况之间的一致程度，以及数据内部、数据之间符合业务逻辑与客观事实的程度。

准确性评价指标及计算方法见表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	数据内容真实性	数据记录反映真实发生的业务活动或客观存在的事实，且具备可靠来源和可追溯性的程度。	$X=A/B$ 式中： X——指标得分；

			A——来源可靠、流程可追溯且经确认反映真实业务活动的数据记录数； B——被评价的事件性数据记录总数。
0302	主体标识唯一性	在同一数据集中，使用指定的主体标识符能够唯一、准确地标识一个经营主体的程度。	$X=A/B$ 式中： X——主体标识唯一性指标得分； A——其标识符在本数据集中唯一且准确指向单一主体的记录数； B——被评价的包含主体标识符的数据记录总数。
0303	权威源比对准确率	关键数据项内容与法定、权威数据源进行比对，结果一致的比例。	$X=A/B$ 式中： X——权威源比对准确率； A——与权威源比对一致的数据项数量； B——应进行权威源比对的数据项总数。
0304	数值范围合理性	数据记录中数值型字段的取值在其业务定义的可能或合理范围内的程度。	$X=A/B$ 式中： X——数值范围合理性指标得分； A——数值落在预设合理值域内的数据项数； B——被评价的数值型数据项总数。 注：合理值域可根据历史数据统计分位数、业务常识或政策规定进行设定。

8.4 一致性评价指标

信用监管数据一致性是指数据与其他特定上下文中，其内容、格式、关系等保持协调一致，无矛盾和冲突的程度。

一致性评价指标及计算方法见表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	数据类型一致性	数据集中数据类型符合其所属数据集类型要求的程度。	$X=A/B$ 式中： X——数据类型一致性指标得分； A——数据类型一致的数据元素个数； B——被评价的数据元素总数。
0402	记录内逻辑一致性	单条数据记录内部，各字段间的取值符合业务逻辑约束的程度。	$X=A/B$ 式中： X——记录内逻辑一致性指标得分； A——通过记录内业务逻辑校验的记录数； B——被评价的记录总数。

0403	关联记录逻辑一致性	具有主外键关联关系的不同记录之间，其状态、数值、日期等符合业务逻辑的程度。	$X=A/B$ 式中： X——关联记录逻辑一致性指标得分； A——关联关系及逻辑正确的关联记录组数； B——应存在关联关系的记录组总数。
0404	跨源数据一致性	描述同一信用主体同一属性的数据，从不同来源或不同业务系统归集后，其核心内容保持一致的程度。	$X=A/B$ 式中： X——跨源数据一致性指标得分； A——跨源比对一致的核心数据项数量； B——应进行跨源比对的核心数据项总数。
0405	时序状态一致性	同一信用主体随时间变化的状态记录，其前后状态转换符合既定规则和时序逻辑的程度。	$X=A/B$ 式中： X——时序状态一致性指标得分； A——状态时序转换正确的记录序列数； B——被评价的状态记录序列总数。

8.5 时效性评价指标

信用监管数据时效性是指数据应符合信息归集公示的时限要求，在时间变化中的正确程度，包括基于时间段的正确性、基于时间点的及时性、时序性、数据授权使用时效性等。

时效性评价指标及计算方法见表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	数据归集及时率	业务事件发生后，在规定时限内（例如，行政许可、行政处罚信息的公示时限要求）被成功归集至信用监管数据系统的数据记录比例。	$X=A/B$ 式中： X——数据归集及时率指标得分； A——在规定时限内完成归集的数据记录数； B——被评价的应按时归集的数据记录总数。
0502	数据更新及时率	当经营主体的核心状态或监管事件状态发生变更后，相关数据项在系统中被更新至最新状态的比例。	$X=A/B$ 式中： X——数据更新及时率指标得分； A——状态变更后，在要求更新周期内完成数据更新的记录数； B——发生状态变更且应更新的数据记录总数。
0503	历史链条连续性	针对特定经营主体的关键监管事件，其所有事件记录在时间轴上被连续、完整归集，无断点或缺失的程度。	$X=A/B$ 式中： X——历史链条连续性指标得分； A——被评价主体中，其特定监管事件链条记录完整、连续的主体数量；

			B——被评价的、应具备该事件链条的主体总数。
0504	数据更新频率	数据集中，数据记录所对应的业务时间或系统时间距离当前评价时点的陈旧程度。	$X = 1 - \bar{T}/T_{\max}$ 式中： X——数据更新频率指标得分； \bar{T} ——被评价数据记录的业务/系统时间距当前时间的平均间隔（如天数）； T_{\max} ——根据业务需求设定的最大可接受陈旧阈值。

8.6 可访问性评价指标

信用监管数据可访问性是指数据在技术层面能够被授权用户或系统有效发现、获取和读取的程度，包括数据可访问性和数据可用性等。

可用性评价指标及计算方法见表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	接口服务可用率	为外部系统或用户提供数据访问的标准化接口在评价周期内处于正常服务状态的比例。	$X = T_{\text{up}}/T_{\text{total}}$ 式中： X——接口服务可用率； T_{up} ——在评价周期内，接口累计正常服务时间； T_{total} ——评价周期的总时间。
0602	数据资源目录可发现性	已归集的数据在统一目录中被有效编目，且元数据完整、准确，便于用户查找和理解的比例。	$X=A/B$ 式中： X——数据资源目录可发现性指标得分； A——编目完整且准确的数据集数量； B——应编目的数据集总数。
0603	授权访问合规性	数据访问行为符合授权范围、权限和期限要求的比例。	$X=A/B$ 式中： X——授权访问合规性指标得分； A——符合授权要求的数据访问记录数； B——被评价的数据访问记录总数。

8.7 AI 可用性指标

信用监管数据AI可用性指标是指数据集在规模、分布、标注、时效等方面，符合用于训练信用监管领域人工智能模型（如风险预警、信用评分、分类画像）特定要求的程度。

AI可用性指标及计算方法见表7。

表7 AI可用性评价指标

指标编号	指标名称	指标描述	计算方法
0701	数据内容多样性	数据集的数据分布全面程度满足目标应用场景人工智能模型开发和训练需求的程度。	$X=A/B$ 式中： X——内容多样性指标得分；

			A——满足内容多样性要求的数据元素个数； B——被评价的数据元素总数。
0702	样本分布均衡性	数据集中,各类别的样本数量分布满足模型训练要求,避免严重类别不平衡的程度。	$X = 1 - A/B$ 式中: X——样本分布均衡性指标得分; A——基尼系数或类别失衡度; B——阈值。
0703	数据标注准确性	数据集中能精准标记出目标应用场景人工智能模型开发和训练所需所有信息的程度。 注:无监督机器学习等不需标注的应用场景不适用。	$X=A/B$ 式中: X——标注准确性指标得分; A——经核查标注正确的样本数; B——被评价的已标注样本总数。
0704	特征数据完备性	对于模型训练所需的特征变量,其在样本中缺失值的比例低于模型可接受阈值的程度。	$X=A/B$ 式中: X——特征数据完备性指标得分; A——缺失率低于阈值的特征项数量; B——模型所需的特征项总数。
0705	时序数据连贯性	用于时序预测模型的数据,其时间序列完整、连续,采样频率符合模型输入要求的程度。	$X=A/B$ 式中: X——时序数据连贯性指标得分; A——时间序列连贯的样本数; B——被评价的时序样本总数。
0706	非结构化数据可处理性	以非结构化形式存储的关键业务文档,能够被有效预处理和特征提取,满足自然语言处理、计算机视觉等模型训练需求的程度。	$X=A/B$ 式中: X——非结构化数据可处理性指标得分; A——可有效处理的非结构化数据样本数; B——被评价的非结构化数据样本总数。
0707	内容原创性	数据集中的内容是新颖的、非复制的。即使数据来源于已知的渠道,但其内容可能是前所未有的组合、特定情境下的记录或是独特现象的描述。	$X=A/B$ 式中: X——内容原创性指标得分; A——经评估为原创的数据样本数; B——被评价的数据样本总数。
0708	内容合规性	数据内容符合国家法律法规、社会主义核心价值观及相关标准规范要求的程度。	$X=A/B$ 式中: X——内容合规性指标得分; A——符合合规要求的数据项数量; B——被评价的数据项总数。

9 评价方法

9.1 评价原则

数据质量评价可采用系统评价、人工评价等方法。宜采用全量、增量以及抽样相结合的策略包括不定期评价、周期性评价以及实时评价方式。

9.2 评价方法

数据质量评价采用定量与定性相结合的方法,具体包括:

- a) 系统自动评价：利用数据质量检测工具对结构化数据进行自动化检测；
 - b) 人工抽样评价：对非结构化数据或复杂业务逻辑进行人工抽样检查；
 - c) 交叉验证评价：通过多源数据比对验证数据一致性。
 - d) 业务规则符合性评价：根据相关业务规范确定的更新周期、覆盖范围等要求进行专项评价。
- 数据质量的最终得分按公式（1）进行计算。一级指标得分按公式(2)计算。

$$P = \sum_{i=1}^7 (w_i \times x_i) \quad (1)$$

公式中：

P ——数据质量最终得分；

i ——第 i 个一级指标， $i = 1, 2, \dots, 7$ ；

w_i ——第 i 个一级指标实际得分；

x_i ——第 i 个一级指标的权重。

$$w_i = \sum_{j=a}^n (s_{ij} \times y_{ij}) \quad (2)$$

公式中：

j ——第 j 个二级指标；

n ——第 i 个一级指标项下的结尾二级指标序号；

a ——第 i 个一级指标项下的起始二级指标序号；

s_{ij} ——第 j 个二级指标（在第 i 个一级指标项下）实际得分；

y_{ij} ——第 j 个二级指标（在第 i 个一级指标项下）的权重。

9.3 指标权重

进行评价时，指标权重宜满足：

- 1) 在面向常规业务应用场景时，评价指标仅包含常规性指标（规范性、完整性、准确性、一致性、时效性、可访问性），六个指标的权重根据应用目的进行划分，每个指标项权重不低于10%；
- 2) 在含有人工智能的应用场景时，常规性指标（规范性、完整性、准确性、一致性、时效性、可访问性）在总评价权重中占比不低于80%，AI可用性指标在总评价权重中占比不高于20%。

参考文献

- [1] 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（ISO/IEC 25024:2015，MOD）
 - [2] GB/T 42755—2023 人工智能 面向机器学习的数据标注规程
 - [3] GB/T 40478—2021 企业信用监管档案数据项规范
 - [4] 《关于开展企业信用监管数据质量全面提升行动的通知》（国市监信发〔2023〕25号）
-