

ICS

CCS

DB61

陕西省地方标准

DB 61/T XXXXX—2022

高速公路信息系统数据处理规程

Data Processing Specifications for Expressway Information System

草案

XXXX - XX - XX 发布

XXXX - XX - XX 实施

陕西省市场监督管理局 发布

目录

目录.....	I
前 言.....	III
高速公路数据处理规程.....	1
1 范围.....	1
2 规范性引用文件.....	1
3 术语和定义.....	1
3.1 数据模型 data model.....	2
3.2 数据清洗 data cleaning.....	2
3.3 数据元 data element.....	2
3.4 元数据 metadata.....	2
4 缩略语.....	2
5 总体要求.....	2
6 数据采集.....	2
7 数据处理过程.....	3
8 数据处理.....	3
8.1 场景解析.....	3
8.1.1 恢复原状型.....	3
8.1.2 防范潜在型.....	3
8.1.3 追求提升型.....	4
8.2 数据探查.....	4
8.2.1 业务探查.....	4
8.2.2 字段探查.....	4
8.2.3 数据集探查.....	5
8.2.4 问题数据探查.....	5
8.2.5 数据应用探查.....	5
8.3 数据定义.....	6
8.3.1 数据格式定义.....	6
8.3.2 数据资源目录.....	6
8.3.3 数据提取策略定义.....	6
8.3.4 数据类目定义.....	7
8.3.5 数据分级分类定义.....	7
8.4 数据清洗.....	8
8.4.1 数据过滤.....	8
8.4.2 数据去重.....	8
8.4.3 格式转换.....	8
8.4.4 数据校验.....	9
8.5 数据提取.....	9
8.5.1 结构化数据提取.....	10
8.5.2 非结构化数据提取.....	10
8.6 数据关联.....	11
8.7 数据分析应用.....	11
8.7.1 描述性分析.....	11

8.7.2 对比分析.....	11
8.7.3 预测性分析.....	12
8.7.4 诊断性分析.....	12
8.7.5 决策性分析.....	12
8.8 数据集标识.....	13
9 数据处理效果评价.....	13
10 数据存储.....	13

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。

本文件由陕西省交通运输厅提出。

本文件由陕西省交通运输标准化技术委员会归口。

本文件起草单位：陕西交通控股集团有限公司、西北工业大学、陕西交通电子工程科技有限公司、西安公路研究院有限公司。

本文件主要起草人：王鑫、田龙、张高峰、苟蛟龙。

本文件由陕西交通控股集团有限公司负责解释。

本文件首次发布。

联系信息：

单位：陕西交通控股集团有限公司

电话：029-87832077

地址：陕西省西安市雁塔区太白南路9号

邮编：710065

高速公路数据处理规程

1 范围

本文件规定了高速公路数据处理全生命周期的技术要求，包括：业务场景解析、数据探查、数据定义、数据清洗、数据提取、数据关联、数据分析应用、数据集标识、数据处理效果评价及数据存储的核心技术要求。

本文件适用于高速公路建设、管理、养护、运营、服务等信息系统数据的集成、预处理、质量评估和存储。

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 31916.3-2018 信息技术 云数据存储和管理 第3部分：分布式文件存储应用接口

GB/T 36344-2018 信息技术 数据质量评价指标

JT/T 697.1-2013 交通信息基础数据元 第1部分：总则

JT/T 697.2-2014 交通信息基础数据元 第2部分：公路信息基础数据元

JT/T 697.9-2016 交通信息基础数据元 第9部分：建设项目信息基础数据元

JT/T 697.13-2009 交通信息基础数据元 第13部分：收费公路信息基础数据元

JT/T 747.1-2020 交通运输信息资源目录体系 第1部分：总体框架

JT/T 747.2-2020 交通运输信息资源目录体系 第2部分：技术要求

JT/T 747.3-2020 交通运输信息资源目录体系 第3部分：核心元数据

JT/T 747.4-2020 交通运输信息资源目录体系 第4部分：公路水路信息资源分类

JT/T 747.5-2020 交通运输信息资源目录体系 第5部分：标识符编码规则

JT/T 747.6-2020 交通运输信息资源目录体系 第6部分：技术管理要求

DB61/T 1896—2024 高速公路数据集成规范

3 术语和定义

下列术语和定义适用于本文件。

3.1 数据模型 data model

按照信息系统中的形式描述和所应用的数据库管理系统的要求构造的数据原型。

[来源：GB/T 5271.17-2010, 3.11]

3.2 数据清洗 data cleaning

数据清洗是指对数据进行重新审查和校验的过程，发现并纠正数据文件中可识别的错误，按照一定的规则把错误或冲突的数据清洗掉，包括检查数据一致性，处理无效值和缺失值等，数据清洗可通过自动化工具或人工干预完成。

3.3 数据元 data element

由一个属性集合规定其定义、标识、表示和允许值的一个数据单元。

[GB/T 18391.3—2009, 3.3.36]

3.4 元数据 metadata

定义和描述其他数据的数据。

[来源：GB/T 18391.1—2009, 3.2.16]

4 缩略语

下列缩略语适用于本文件。

URL：统一资源定位符（Uniform Resource Locator）

OCR：光学字符识别（Optical Character Recognition）

NLP：自然语言处理（Natural Language Processing）

5 总体要求

数据处理的总体要求如下：

1) 数据处理应按照标准化、规范化的处理流程，流程包括但不限于：场景解析、数据探查、数据定义、数据清洗、数据提取、数据关联、数据分析应用、数据集标识、数据处理效果评价和数据存储10个步骤。

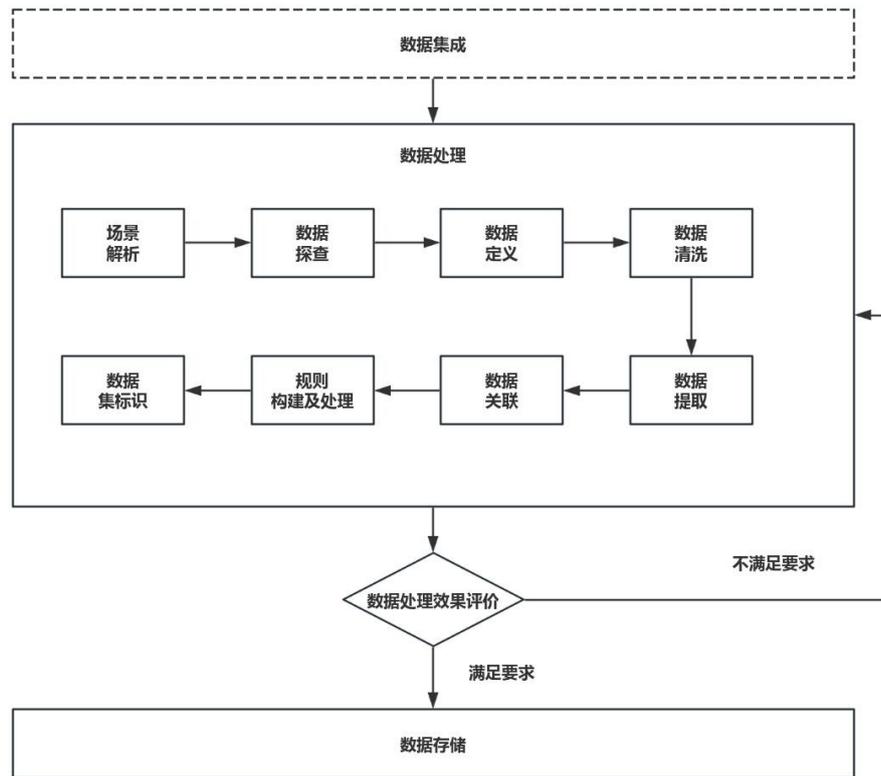
2) 数据处理中数据安全应符合GB/T 22239中8.1的规定。

3) 数据处理应明确业务场景，并满足业务场景对性能的要求。

6 数据采集

数据采集参考《高速公路数据集成规范》（DB61/T 1896-2024）

7 数据处理过程



数据处理过程图

8 数据处理

8.1 场景解析

场景解析应结合高速公路相关的业务现状和预期达到目标，定义数据处理的场景。包括但不限于以下三种场景：

8.1.1 恢复原状型

此类场景解析应明确业务上暴露出的问题具体情况，目的是通过数据处理手段恢复原状，需包括出现问题的对象描述、时间、现象、影响因素。

8.1.2 防范潜在型

此类场景解析应包括业务上搁置不管会发生哪些不良状态,目的是通过数据处理后辅助决策维持业务现状,不良状况描述应包括涉及哪些对象、搁置风险周期、潜在风险、可能影响四要素。

8.1.3 追求提升型

此类场景解析需明确对象、现状基线、提升目标、增效路径四要素。其中对象需明确提升的交通主体,现状基线应明确当前交通主体的关键指标,关键指标包括但不限于转换率、故障率;提升目标应明确交通主体提升的量化目标和时间节点;增效路径应明确实现目标的具体策略、工具与执行步骤。

8.2 数据探查

数据探查应通过规则对数据进行探测,从数据业务含义、字段格式语义、数据结构、数据质量等多个维度进行探查,以达到认识数据的目的,为数据定义提供依据。

8.2.1 业务探查

业务探查应获取数据来源单位、所属应用系统、业务含义描述、安全性要求、主外键名称、表关联关系等内容。

业务探查应利用智能分析技术或人工分析的方式,对来源数据集与已有的业务数据集进行相似性分析,通过相似度结果来确定数据集的业务含义。

8.2.2 字段探查

8.2.2.1 空值率探查

空值率探查应对每个字段内容是否为空进行探查,支持探查字段的动态配置,字段空值率的计算方法:

(物理表字段空值行总数/物理表字段总行数)*100%。

空值率探查应依据元数据标注识别允许为空的特殊字段,并记录例外日志供审计复核。通过字段空值率探查,需实现对数据质量的初步评估,一方面可对空值率较高字段进行重点关注,反馈给数据提供方进行数据排查,追踪上游数据情况;另一方面与字段历史空值率进行比对,及时发现数据质量的动态变化。

8.2.2.2 值域及分布探查

值域探查应包括但不限于:

a) 字典值探查:通过探查掌握数据项的字典表信息。示例:高速公路病害程度的字典表是“1-严重;2-中等;3-轻微。”

b) 取值范围探查：根据字段内容进行取值范围探查，确定最大值、最小值。示例：地球经度的取值范围是在“-180至180度之间”。

数据分布探查应统计数据项每个值的具体分布情况，供业务分析使用。

8.2.2.3 类型及格式探查

应对字段数据类型和数据格式进行探查，形成的探查结果为数据定义提供输入。

数据类型探查应探查字段内容存储所采用的数据类型。示例：日期字段的数据类型。

数据格式探查应探查字段内容的格式和数据长度。示例：车牌号的数据格式和数据长度。

8.2.2.4 命名实体探查

应通过探查识别字段内容中身份证号、统一社会信用代码、车牌号、手机号码、电话号码、邮政编码、邮箱、经度、纬度、主体身份代码、码表、日期、日期时间、组织机构代码等命名实体，为数据处理中数据清洗环节提供数据处理依据。敏感数据探查应严格限定授权人员操作，探查结果应加密存储，禁止直接暴露原始敏感信息。

应通过规则化、智能化方法对字段内容进行分析，根据分析结果对字段内容进行标识。

8.2.3 数据集探查

应通过对数据集表名称的文本语义分析，与标准数据集名称进行对比分析，探查数据集与标准数据集的匹配情况。

8.2.4 问题数据探查

问题数据探查应探查字段中不合理的信息，为后续数据清洗规则的制定提供依据。探查问题分类包括但不限于代码字典表问题、数据类型问题、数据值逻辑问题、数据格式问题、必填项为空等。

应对每次探查结果记录并形成报告。

问题数据产生原因包括但不限于：

- 1) 操作方面：因业务人员疏忽在数据录入阶段产生的问题数据；
- 2) 技术方面：因技术问题导致在数据传输过程中发生异常产生的问题数据；
- 3) 理解方面：因对数据表业务理解或生成逻辑理解不清，导致在数据处理过程中产生的问题数据。

8.2.5 数据应用探查

应明确数据应用对方的数据要求和数据应用场景范围，结合数据分级分类策略建立细粒度授权规则，防范超范围使用风险。

8.3 数据定义

数据定义应根据数据探查结果，结合业务需求，定义数据处理内容和方法。

数据定义结果应随着数据探查结果和业务需求的变更而动态维护，以元数据的形式描述输出到元数据库中。

数据定义包括但不限于：数据格式定义、数据类目定义、资源目录注册、数据分级分类定义、数据清洗策略定义、数据关联策略定义、数据比对策略定义、数据标识策略定义、数据质量核检规则定义等。

8.3.1 数据格式定义

应根据数据探查中业务探查和字段探查的结果，建立源数据中原始字段项与标准数据元的映射关系，以及原始字典代码集与规范化字典代码集的映射关系。

1) 原始字段项与标准数据元映射关系应满足：

通过数据元对标操作，明确源数据中各原始字段项与《陕西省交通信息基础数据元字典（公路）》中标准数据元的对应关系。如果出现新的数据元，应进行注册与登记。

2) 原始字典代码集与规范化字典代码集映射关系的定义：

通过代码集对标操作，明确源数据使用的原始字典代码集（值）与标准数据元使用的规范化字典代码集（值）之间的对应关系。

8.3.2 数据资源目录

数据资源目录应按照统一的陕西省高速公路数据资源目录标准规范，对高速公路数据资源进行统一管理，实现数据资源科学、有序、安全使用。数据资源目录定义应包含数据资源信息数据项和数据项信息数据项两部分内容定义。

1) 数据资源信息应包含但不限于：数据资源事权单位、数据资源目录编号、数据资源标识符、数据资源中文名称、数据资源英文名称、数据资源描述、数据资源类别、数据资源更新周期、数据资源更新方式、存量数据记录规模、存量数据存储规模、数据资源存储周期、数据资源状态、注册时间、更新时间、数据资源共享类型、数据资源共享条件、数据资源共享方式类型等。

2) 数据项信息应包含但不限于数据项编号、数据项标识符、中文名称、英文名称、关联数据元标识符、限定词、说明描述、数据项长度、数据项类型、值域、计量单位、是否标识项、是否必填、是否分词、是否查询条件、数据项敏感程度、注册时间、更新时间、版本、是否失效等。

8.3.3 数据提取策略定义

需按照业务需求，定义从来源数据提取所需数据的策略。来源数据包括原始库、资源库、主题库和专题库等。

1) 结构化/半结构化数据提取策略的定义：

针对结构化/半结构化数据，根据数据组织或业务需求，明确来源数据资源/数据项到目的数据资源/数据项的提取映射关系。

2) 非结构化数据提取策略的定义：

需针对非结构化数据，根据业务需求，描述从非结构化数据中提取要素及其相互关系、事件信息过程等策略，包括但不限于以下内容：

a) 来源数据描述：通过数据资源标识符，明确数据提取策略所对应的来源数据资源；

b) 来源数据格式：描述来源数据资源的格式类型，例如文本、图像、音频、视频等；

c) 目的数据描述：通过数据资源标识符和数据项标识符，明确数据提取策略所对应的目的数据资源/数据项；

d) 提取策略描述：描述非结构化数据的提取策略，包括要素提取、摘要提取、特征提取、声纹提取、语种提取、语音识别、视频片段提取等。

8.3.4 数据类目定义

对于接入到数据资源池的数据应定义数据类目来组织接入的数据，接入的数据类目要遵循一定的规范，如按数据来源部门进行类目定义、按数据结构方式进行类目定义、按数据所属业务域进行类目定义等。

要求数据类目和类目下数据解耦，类目可自由定义、灵活调整，方便数据组织和数据层次结构的划分。

要求类目有一定的元数据信息，如类目的中文名称、类目简称等，方便在数据处理过程中自动归集数据，实现数据归集智能化。

要求类目管理权限的控制，仅限有权限的人调整类目，类目层级可自由调整。

8.3.5 数据分级分类定义

数据分级分类应通过描述数据的多维度特征和内容敏感程度，为定制数据资源的开放和共享策略提供支撑。根据数据内容的敏感程度对数据资源进行定级，按照数据级别控制数据资源的使用范围。从数据资源种类、数据项敏感度等多个维度对数据资源进行分类，按照数据类别控制数据资源的使用范围。

具体分级分类定义方法参见《陕西省公路数据资源分类分级规范》。

8.4 数据清洗

数据清洗应根据数据定义结果进行数据过滤、去重、格转、校验等操作，包括流式数据清洗和批量数据清洗，生成满足质量要求的数据，并存入资源库中。

清洗过程中敏感数据的保护机制：

- 1) 数据清洗需在隔离环境中进行，禁止明文传输敏感数据；
- 2) 清洗过程中应对个人信息（如身份证号、手机号）进行脱敏处理；
- 3) 清洗日志应记录操作人员、时间及操作内容，并定期审计。

8.4.1 数据过滤

应通过对数据进行辨别和分离，实现冗余数据及垃圾数据的滤除，被识别为垃圾数据的数据可以直接滤除，或标识后正常处理并交由后端模块判断如何进一步处理。其他未被过滤掉的数据则直接进入后续处理环节。

1) 基于样本数据的垃圾数据过滤：根据数据定义的过滤对象，调用垃圾样本库，对例如互联网等来源数据进行辨别、分离和过滤；

2) 基于规则的垃圾数据过滤：系统根据过滤规则对数据进行过滤。垃圾数据的处理应支持垃圾数据单独存储；应支持对垃圾数据进行标识；应支持垃圾数据在全文数据中的检索。

8.4.2 数据去重

在各类场景下需设定相应的数据重复判别规则以及合并、清除策略，对数据进行重复性辨别，并对重复数据进行合并或清除处理。根据数据定义环节定义好的重复规则判定结构化数据是否重复，如果数据重复，则直接滤除或者进入问题库。

8.4.3 格式转换

来源数据可能存在多种不同的格式，需要根据数据元标准把非标准数据转换成统一的标准格式进行输出，将不同来源的同类数据按照统一规则进行转换。

格式转换内容包括但不限于：

- 1) 代码转换：将不同来源数据中的用代码表示的数据，统一转换成标准的代码规范内容；
- 2) 数据内容格式统一：对车牌号、身份证号、统一社会信用代码、手机号码、电话号码、邮政编码、邮箱、经度、纬度、主体身份代码、码表、日期、日期时间、组织机构代码等格式统一，格式统一后的字段单独存储。包括但不限于以下数据内容格式统一：
 - a) 特定字段全角转半角：手机号码、URL、账号等；

b) 经纬度：将各种类型的坐标系统一转换为“2000国家大地坐标系”（CGCS2000）并保留原始数据及字段信息，经纬度坐标值单位采用“十进制度”表示，定位精度须至少精确到小数点后5位（即与实际位置偏差在10米范围内），正值表示东经、北纬，负值表示西经、南纬，取值范围经度-180.00000~180.00000，纬度-90.00000~90.00000。例如，123.23000表示东经123.23000度；-133.00000表示西经133.00000度。

8.4.4 数据校验

应根据数据质量检核规则对数据进行检验，符合标准的数据直接入库，不符合标准的数据可进入问题数据库以便进一步分析处理。校验主要包括数据的完整性校验、一致性校验等。常用的校验规则有：空值校验、取值范围校验、公民身份证号码等校验、数值校验、长度校验等。此外，还有更为复杂的多字段条件校验、业务规则校验等。

校验内容包括但不限于：

- 1) 空值校验：对数据字段的空值情况进行校验，必填项进行补充或进入问题库；
- 2) 取值范围校验：对数据字段的取值范围进行校验，不符合范围的数据进入问题库；
- 3) 居民身份证号码/手机号等校验：对数据字段格式和内容进行校验，不符合数据检核规则的数据，进入问题库或另行处理。比如，公民身份证号码可按照6位数字行政区划码+8位数字出生日期码+3位数字顺序码+1位校验码的编码规则进行精准校验，也可进行末位不检验的粗略校验；
- 4) 数值校验：对数值型数据字段进行校验，不符合数值型，进行格式转换并存储，无法转换成数值型的数据进入问题库；
- 5) 长度校验：对数据字段的长度进行校验，不符合长度的数据进行截取，进入问题库；
- 6) 多字段条件校验：对多个相关字段进行互相校验，字段间出现矛盾的数据，进入问题库；
- 7) 业务规则校验：按照一定的业务规则进行校验，不符合业务规则的数据，进入问题库。

8.5 数据提取

数据提取应根据数据定义，从源格式数据中提取出目的格式数据，根据不同的提取策略将提取结果存入资源库、主题库、专题库、知识库等逻辑数据库。数据提取应支持组件化，可扩展，可配置。

8.5.1 结构化数据提取

结构化提取的来源和目的数据格式均为结构化，应根据数据组织或业务需求进行数据的转换及整合，获得按照目的数据形式组织的数据。结构化提取首先获得结构化提取策略或规则并进行解析，得到从来源数据集/字段到目的数据集/字段的映射关系、运算规则等，然后按照规则实施结构化提取。

8.5.2 非结构化数据提取

非结构化数据包括但不限于办公文档、网页、文本、图像、视频等。这些数据需进行结构化提取才便于进行进一步的计算和使用。

8.5.2.1 文本数据提取

通过自然语言处理技术，应从文本数据中提取姓名、身份证号、手机号、统一社会信用代码等要素及其相互关系、事件等信息。待提取的文本数据可以是网页、电子邮件、即时通讯聊天信息、办公文档和各类报表（如wps、pdf、doc、xlsx）。

文本数据提取内容包含但不限于以下内容：

1) 要素提取：应利用OCR、正则表达式、NLP实体识别方法从文本中提取出各类通用要素或者业务要素信息。如：车牌号、公民身份证号、手机号、统一社会信用代码等；

2) 关键词和摘要提取：应利用关键词提取法、深度学习法、抽取式摘要法从文本中提取关键词和关键段落构成摘要内容，方便用户快速预览文本内容；

3) 关系提取：从文本中提取出要素之间的关系。比如人与车辆之间的所属关系。

8.5.2.2 视频、图像数据提取

应从视频、图像数据中提取文字、图标、车辆情况等实体信息，以及相关特征信息。视频、图像数据提取内容包含但不限于以下内容：

1) 视频、图像对象特征提取：对视频、图像数据对象特征的提取内容包含：视频片段对象、图像对象、车辆对象、场景对象、视频事件对象、视频图像标签对象；

2) 视频元数据内容提取：视频元数据内容提取，包括制作人姓名、创建时间、修改时间、文件大小、文件类型、文件名称、编码、帧率、平均帧率、起始时间、结束时间、时长、码率、帧数等；

3) 图像元数据内容提取：图像元数据内容提取，包括创建时间、修改时间、文件大小、文件类型、文件名称、尺度、高度（像素）、宽度（像素）、水平分辨率（dpi）、垂直分辨率（dpi）、颜色、位置、设备类型等。

8.6 数据关联

数据关联应将数据和其他业务数据等进行关联，并输出关联信息，关联回填和关联提取的结果存入资源库或主题库中。

数据关联的主要功能包括以下：

1) 关联回填：通常是将数据资源与其他业务数据进行关联，并将关联的信息回填到日志，提升数据的关联及价值；如车辆关联：根据数据内容回填车辆的命名实体标识信息；

2) 关联提取：根据主题数据定义，对各类数据资源中有业务关联关系的数据进行提取。

8.7 数据分析应用

围绕关联后高速公路数据开展分析，应至少从描述性分析、对比分析、预测性分析、诊断性分析、决策性分析等角度开展。

8.7.1 描述性分析

描述性分析应对数据进行总结、描述和可视化，揭示高速数据的基本特征、分布和趋势。

描述性分析的主要包括但不限于：

1) 统计指标。如均值、中位数、标准差、最大值、最小值等，来描述数据的集中趋势、离散程度和分布形态。

2) 频数分布。展示数据分布情况的方法，将数据按照数值或类别进行分组，并计算每个组的频数。

3) 可视化。通过可视化方法包括折线图、散点图、箱线图、热图等发现高速数据中的模式、趋势和异常值，提供直观的数据分析结果。

4) 变异系数。衡量数据变异程度的方法，是标准差与均值的比值，用于描述不同变量或不同群体之间的变异程度。

5) 相关分析。衡量两个变量之间相关关系的方法，取值范围是-1到1，接近1表示正相关，-1表示负相关，接近0表示无关。

8.7.2 对比分析

数据比对应按照规则对数据进行相同比较或相似度计算, 对于命中规则的数据，按照输出描述进行输出，并对命中结果数据进行存储。

1) 结构化数据比对

应将比对目标与比对源指定字段的取值进行比对，实时发现比中信息。支持完全匹配、模糊匹配、范围匹配、正则匹配等。

完全匹配：检索比对目标内容与比对源字段内容完全相同；

模糊匹配：比对目标内容在比对源字段内容中出现，则匹配成功；

范围匹配：比对目标内容与比对源指定的字段进行对比，指定的字段内容在比对目标内容区间，则匹配成功。比如某个坐标系范围内的比对；

正则匹配：比对目标内容为正则表达式，比对源指定的字段内容符合比对目标内容设定的规则，则匹配成功。

2) 非结构化数据比对

应将比对目标与非结构化数据比对，在非结构化数据中实时发现比对目标相关信息。支持关键词比对、文本相似度比对、二进制比对等。

关键词比对：用户可设定关键词规则，对接收到的非结构化数据进行内容比对。关键词规则需支持逻辑表达式最大可支持64个32字节长的关键词组合。

文本相似度比对：对目标文本进行特征抽取，与比对源中的文本数据进行文本特征比对，返回相似度数值结果、对应文本信息。

二进制比对：对二进制比对目标文件进行MD5值计算，与非结构化数据的MD5值进行比对，返回相似度数值结果、对应非结构化数据。

8.7.3 预测性分析

预测性分析应利用历史数据和统计模型来预测未来事件或者趋势，包括但不限于回归分析、时间序列分析及预测、人工神经网络、支持向量机，并最小化预测值与真实值的误差。

8.7.4 诊断性分析

诊断性分析应利用系统性方法用于识别问题的原因、评估问题的影响，并提供解决方案。诊断性分析的方法包括但不限于根本原因分析、逻辑推理、实证研究，并通过实证研究和实验，收集和分析数据，以验证问题的原因和解决方案的有效性。

8.7.5 决策性分析

决策性分析应基于最小化原则，通过收集、整理和分析相关数据和信息，以支持决策过程中的选择和决策制定。决策性分析的方法包括但不限于以下几种：

1) 统计分析：通过收集和分析大量的数据，运用统计学原理和方法，来揭示数据之间的关系和趋势。常用的统计分析方法包括描述统计、推断统计、回归分析、方差分析等。

2) 决策树分析：将决策问题以树状结构进行表示，通过对不同决策路径的评估和比较确定最优的决策方案。决策树分析方法可以帮助决策者厘清决策过程中的关键因素和可能的结果。

3) 敏感性分析:通过对决策模型中的关键变量进行变动和调整,评估这些变动对决策结果的影响。敏感性分析可以帮助决策者了解决策结果的稳定性和可靠性,并做出相应的调整和决策。

4) 成本效益分析:通过比较决策方案的成本和效益,评估不同决策方案的经济效益和决策可行性。成本效益分析可以帮助决策者在有限资源下做出最优的决策。

5) 多属性决策分析:将决策问题中的多个属性或指标进行量化和评估,通过综合考虑这些属性的权重和重要性,确定最优的决策方案。常用的多属性决策分析方法包括层次分析。

6) 模拟分析:通过建立决策模型和模拟实验,模拟不同决策方案的结果和影响。模拟法、模糊综合评价法等。分析可以帮助决策者在实际决策前进行预测和评估,以减少决策风险。

7) 基于专家判断的分析:依靠领域专家的经验 and 知识,通过专家评估、专家意见征询等方式,对决策问题进行分析 and 判断。专家判断可以提供有关决策问题的专业见解 and 建议。

8.8 数据集标识

通过上述数据处理过程形成的原始数据集、处理后的高质量数据集应基于标签知识库、标签引擎对数据集进行标识,对资源库、主题库、专题库、知识库的数据资源打上标签,为上层应用提供支撑。

9 数据处理效果评价

数据处理效果评价应遵循DB 61/T 1896-2024相关规定,评估指标包括唯一性和安全性。

唯一性主要用于度量与评估数据资产内容或相关属性的重复情况,包括主键唯一性和数据唯一性。安全性主要用于描述数据在传输、存储和访问过程中的安全特性,以确保数据不被未经授权的访问、篡改或泄露。包括传输安全性和存储安全性。

10 数据存储

应按照GB/T 31916.3-2018的规定执行。数据处理的结果数据应按照各自的属性及业务要求存储至对应的存储位置,且能满足高性能、高吞吐率、大容量、安全的数据存储条件。数据存储应符合国家及行业相关的安全标准,包括但不限于数据加密、访问控制、备份恢复等,以确保数据的安全性和可用性。

应存储在集中化的存储仓库,存储仓库对各种类型数据库有机融合提供统一访问端口,包括关系型数据库、键值存储数据库、列存储数据库、文档数据库、图形数据库及文件系统等,可根据需求选择或组合相应的数据存储方式。

参考文献

- [1]陕西省公路数据质量评估规范
- [2]陕西省公路行业数据采集更新管理办法20211112
- [3]陕西省交通运输厅《陕西省交通运输政务信息资源目录（2018版）》
- [4]陕西省公路局《陕西省交通信息基础数据元字典（公路）》
- [5]陕西省公路局《陕西省公路数据资源分类分级规范》