乐山市政务数据清洗加工规范

# 说明

## 目的

为推进全市政务数据资源归集共享，提升政务数据质量，促进数据要素转换，为数据开发利用提供支撑，特制定本规范。

## 内容概述

本规范主要包含政务数据清洗加工流程、数据抽取、规则定义、数据过滤、数据核验、错误标识、修正处理、数据转换、结果检验、数据回流。清洗加工后的政务数据内容需符合一致性、完整性、规范性、准确性、及时性要求，保证共享开放数据能满足数据需求方的应用需求。

## 术语和定义

1.数据清洗加工

运用一定方法修正识别到的数据问题，提升数据质量的过程。

2.政务数据

各级政府部门及其技术支撑单位在履行职责过程中采集、生产、存储、管理的各类数据资源。根据可传播范围，政务数据一般包含可共享政务数据、可开放公共数据和不宜开放共享政务数据。

1.噪声数据

无意义数据，或所有难以被机器正确理解和翻译的数据。

2.结构化数据

一种数据表现形式，按此种形式，由数据元素汇集而成的每个记录的结构都是一致的并且可以使用关系模型予以有效描述。

3.非结构化数据

不具有预定义模型或未以预定义方式组织的数据。

4.半结构化数据

具有结构性，但结构变化大，且难以用结构化数据的处理方法将其放进二维表的数据。

5.表结构

为主体层内容提供表示语义的一种存储范例。

6.ETL

数据的抽取、转换和加载（Extract Transform Load）。

7.ELT

数据的抽取、加载和转换（Extract Load Transform）。

## 适用对象

本规范适用于共享交换平台归集到的所有政务数据，指政务部门在履行职责过程中产生或获取的，以一定形式记录、保存的文件、资料、图表和数据等各类信息资源数据，包括政务部门直接或通过第三方依法采集的、依法授权管理的和因履行职责需要依托政务信息系统形成的信息资源数据等。

## 其他说明

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。

本文件由乐山市数字经济发展局提出。

# 政务数据清洗加工规范

## 清洗加工流程

## 数据抽取

1.抽取来源共享交换平台，通过前置库等方式进行抽取。

2.应具备全量抽取和增量抽取两种方式。

3.应支持结构、半结构和非结构等不同类型数据的抽取。

4.抽取目的地的存储容量应能支持抽取来源的数据总量，数据抽取目的地的表结构应与抽取来源的表结构保持一致。

5.增量抽取（更新）应确定增量更新的方式，抽取的数据应有字段可区分，如更新时间等。

## 定义规则

1.应分析抽取目标数据的范围、体量、类型、内容、关系、质量等信息，全面认识数据情况。

2.以需求为导向、应用为目标，考虑目标数据资源特点和工作复杂程度，结合业务要求或用户和其他相关方的需求、期望，确定切实可操作的数据清洗加工日标及规则。

## 数据过滤

1.将非结构化和半结构化数据转化为结构化数据；

2.对噪声数据进行删除；

3.对业务数据中不符合应用规则的数据进行删除；

4.过滤删除掉的数据应存入问题数据库表，便于后续查证或重新使用。

## 数据检验

1.检核要求

（1）非空检核

应在字段为非空的情况下，对该字段数据进行检核，数据不能为空值。

（2）长度检核

数据长度应满足转换要求的字段长度。

（3）数据量检核

过滤后的数据总量应与原始抽取的数据总量吻合。

（4）数据类型和值检核

数据类型和值应能支持后续数据转换过程，如后续根据定义规则需要将时间字符串数据转换成时间类型时，还需检验时间字符串类型的数据，应符合时间格式。

2.检核步骤

数据检核步骤如下：

（1）按检核要求规定检核数据是否满足要求；

（2）当不满足检核要求时，应进行数据错误标识；

（3）当满足检核要求是，直接进行数据转换。

## 错误标识

1.错误类型

常见数据的错误类型如下：

（1）残缺数据：缺一些记录，或一条记录里缺一些值（空值），或两者都缺；

（2）错误数据：数据没有严格按照规范记录，包括格式内容错误、逻辑错误、不合规等；

（3）重复数据：相同的记录出现多条或多条记录代表同一实体。

2.识别方法

可采用统计学方法、数据挖掘、基于聚类的方法、基于距离的方法、基于分类的方法、基于关联规则的方法、业务区分等方式分析数据，从而识别出数据的错误类型。

3.标识步骤

错误标识步骤如下：

（1）按上文识别方法中推荐的方法，分析筛选出数据资源中存在的数据问题；

（2）按上文错误类型中的类型，对数据问题进行分类，标识错误。

## 修正处理

1.残缺数据处理

（1）处理策略

按照字段缺失比例和字段重要性，分别制定策略。残缺数据处理策略制定参考图见图3。



图3 残缺数据处理策略制定参考图

（2）去除字段

备份当前数据，直接删掉不需要的字段。

（3）填充确实内容

按照以下方式填充确实内容：

a)不同指标的计算结果填充：通过数据项与数据项之间的逻辑联系，采取一定的列拆分、列计算等方式得到缺失内容，如年龄字段缺失，但具有公民身份证号，可从公民身份证号提取年龄数据；

b)同一指标的计算结果填充：采取均值、中位数、众数等方式进行填充，如时间序列缺失，可使用前后的均值填充；

c)重新获取：当缺失率高且非常重要的数据项，应采取重新抽取不同数据源的数据进行关联对比填充。

2.错误数据处理

（1）格式内容问题

格式内容问题数据处理采用以下方法：

a)全、半角处理：通过正则表达式将全、半角符号按照事先定义的规则进行全、半角符号统一；

b)有不该存在的字符：以半自动校验结合半人工方式来找出存在的问题，自动去除不需要的字符，将数据自动化统一或人工修正为正确字符；

c)内容与字段不匹配：详细识别问题类型，如人工填写错误、前端没有校验、导入数据时部分或全部存在列没有对齐、数据源端业务系统缺陷等，不能直接删除，应按照清洗规则，采取加入更多数据源进行数据关联，找到匹配的相应字段进行填补。

（2）逻辑问题

逻辑问题数据处理采用以下方法：

a)了解数据潜在的逻辑规则，采取逻辑推理法，直接去掉一些使用简单逻辑推理即可发现问题的数据；

b)对于不重要的不合理数据宜直接删除，对于重要的不合理值宜进行人工干预或引入更多数据源进行关联识别；

c)通过字段间相互验证的方法修正矛盾内容，如根据字段的数据来源，判定哪个字段提供的信息更可靠，去除或重构不可靠字段；

d)通过分箱、聚类、回归等方法识别离群值（异常值），按照经验和业务流程判断其合理性，若合理，则保留该数值；若不合理，对重要性较高而无法重新采集的数值，按缺失数据处理，对重要性较低的数值，可直接删除；

e)对于复杂逻辑数据问题应咨询了解该数据的产生原因，按照协商的清洗加工规则进行处理。

（3）不合规问题

不合规问题数据处理采用以下方法：

a)设定判定规则：设定强制合规条件，不在规则范围内的，强制设置最大值及最小值，剔除或判断为无效字段；

b)设定警告规则：不在规则范围内的，进行警告及人工处理。

3.重复数据

重复问题处理步骤如下：

通过元数据血缘关系查询到重复数据的各个来源；

通过数据主键或寻找相关信息识别重复数据的含义，不是相同含义的数据不能界定为重复数据今昔去重处理，应分别保留；

查询到确定的重复数据，根据权威性和应用场合，选择最恰当渠道来源的数据，或在不影响数据保真度和完整性的情况下进行合并处理。

## 数据转换

1.应在数据检验通过后开始。

2.开始前应检查需要转换的数据规则和字段是否一致。

3.应实现对数据的格式、信息代码、值的冲突进行转换，典型业务数据转换规则见附录A。

4.转换后的数据结构应与目标数据库的结构相兼容。

5.数据向目标移动时，将其从源数据中移除，或数据复制到多个目标中。

6.转换失败应立即停止，开始查找问题。

7.长时间未转换结束，需仔细核查数据量、规则和字段是否一致，如有问题应立即停止。

8.转换中查找到问题，应解决问题后再开始数据转换。

## 结果检验

1.检核内容

（1）主键重复

检验多个业务系统中同类数据经过清洗后，在统一保存时，主键的唯一性。

（2）非法代码、非法值

检查个别字段出现的异常信息，包括非法代码、代码与数据标准不一致、取值错误、格式错误、多余字符、乱码等。

（3）数据格式

检验表中属性值的格式正确衡量其准确性，如时间格式、币种格式、多余字符、乱码。

（4）记录数

检验各个系统相关数据之间的数据总数检核或数据表中每日数据量的波动。

（5）业务约束

应从业务的角度考虑数据的正确性、一致性、有效性等，如建档日期、入学日期、民族信息等。

（6）约束标准

对照系统数据应符合的标准进行核对。

2.结果要求

（1）规范性

数据的质量及存储标准应统一，源数据应在源头或备份表中能找到。

（2）重复性

数据在字段、记录内容或数据集内不应有重复值。

（3）准确性

数据所指内容对数据所指对象的反应、表现应准确，数据形式对数据内容的表述、表达应准确。

（4）完整性

数据集合中应包含足够的数据来响应各种查询和支持各种计算。

（5）一致性

数据的一致性应符合以下要求：

a)同一个数据在同一时刻在不同数据库、应用和系统中应只有一个值；

b)数据字段内数据应与字段描述一致；

c)最终结果数据的统计量应与预测一致。

（6）时效性

不同类型的应用对数据的时间特性有不同的要求，数据的时间特性应满足业务应用的要求，数据记录应根据时间特性及时更新。

（7）稳定性

数据来源稳定，数据结果能支撑后续业务。

3.检验步骤

结果检验包括一下步骤：

（1）按检验内容规定检验清洗加工后的数据资源情况；

（2）按检验要求核对数据资源达标情况；

（3）当数据资源未达到检验要求，应返回再次进行数据转换；

（4）当数据资源达到检验要求，应进行数据加载或结束清洗加工。

## （十） 数据回流

数据清洗加工后通过库表、接口等方式回流到共享交换平台，以供数据共享开放得高质量应用提供标准化数据支撑。

**附录 A**

**（资料性）**

**典型业务数据转换规则表**

|  |  |
| --- | --- |
| 转换规则 | 规则描述 |
| 统一时间日期数据格式 | 将各类日期统一为八位的字符日期，如YYYYMMDD； |
| 将各类时间统一为六位的字符时间，如HHMMSS； |
| 将各类事件日期统一为十四位的字符时间日期，如YYYYMMDDHHMMSS； |
| 统一分类数据取值代码 | 将人员的性别数据统一转换为国标性别信息代码； |
| 将人员的民族数据统一转换为标准信息代码； |
| 将人员的户籍地址数据统一转换为行政区划代码； |
| 将人员的婚姻登记情况统一转换为标准的婚姻状况代码； |
| 将公民身份证统一转换为18位的身份证号。 |