ICS XXXXXXX
CCS X XX

DB65

新 疆 维 吾 尔 自 治 区 地 方 标 准

DB65/T XXXX-2024

一体化数据资源体系
公共数据采集规范
（征求意见稿）

2024-XX-XX 发布　　2024-XX-XX 实施

新疆维吾尔自治区市场监督管理局　　发 布

目　次

前　言

本标准按照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》的规则编制。

本标准由新疆维吾尔自治区数字化发展局提出并归口。

本标准主要起草单位：新疆维吾尔自治区数字化发展局。

本标准主要起草人：

一体化数据资源体系公共数据采集规范

1 范围

本文件规定了公共数据采集的术语与定义、数据采集要求、数据采集方案、数据采集内容、数据映射关系、数据采集方式、数据采集过程、绩效评价等相关内容。

本文件适用于公共管理和服务机构在履行职责和提供公共服务过程中所涉及的数据采集活动。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是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 35274-2023 信息安全技术 大数据服务安全能力要求

GB/T 36625.3-2021 智慧城市 数据融合 第3部分：数据采集规范

GB/T 22239-2019 信息安全技术 网络安全等级保护基本要求

GB/T 36344-2018 信息技术 数据质量评价指标

GB/T 35295-2017 信息技术 大数据 术语

3 术语与定义

下列术语和定义适用于本文件。

3.1

公共数据 public data

包括政务数据和公共服务数据。政务数据，是指国家机关和法律、法规授权的具有管理公共事务职能的组织（以下称政务部门）为履行法定职责收集、产生的数据。公共服务数据，是指医疗、教育、供水、供电、供气、通信、交通、文旅、体育等公共企业事业单位（以下称公共服务部门）在提供公共服务过程中收集、产生的涉及公共利益的数据。根据应用需求，税务、海关、金融监督管理等国家有关部门派驻自治区管理机构提供的数据属于公共数据。

3.2

数据采集 data acquisition

从数据源中得到原始数据，通过标准化处理并转化为满足数据共享与利用需求的过程。

[来源：GB/T 36625.3—2021，3.2]

3.3

一体化数据资源服务平台 integrated data resource service platform

自治区数字化发展过程中提供数据归集、数据共享、数据开放、数据计算等服务的基础平台。

3.4

数据提供方 data provider

通过一体化数据资源服务平台提供数据资源的政务部门和公共服务部门。

3.5

结构化数据 structured data

一种数据表示形式，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。

[来源：GB/T 35295-2017，2.2.13]

3.6

半结构化数据 semi-structured data

一种结构化数据表示形式,它并不符合关系型数据库或其他数据表的形式关联起来的数据模型结构，但包含相关标记，用来分隔语义元素以及对记录和字段进行分层。

3.7

非结构化数据 unstructured data

不具备预定义模型或未以预定义方式组织的数据。

[来源：GB/T 35295-2017，2.1.25]

4 数据采集要求

4.1 通用要求

4.1.1 数据采集的种类和范围与其履行的公共管理职责或者提供的公共服务范围相适应，且为依法提供公共服务或者履行公共管理职责所必需。

4.1.2 能够通过共享方式获得的数据，不得以其他方式重复采集。

4.1.3 各级政务部门和公共服务部门应对所采集的公共数据资源进行统一编码，自然人数据以有效身份号码作为标识进行采集，法人以及非法人组织数据以统一社会信用代码作为标识进行采集，自然资源和空间地理数据以地理编码作为标识进行采集。

4.1.4 一体化数据资源服务平台方应建立统一的采集管理制度，满足一体化数据资源体系总体框架的要求，并配备系统管理员，对数据提供方实行逐级授权管理，实现管理角色的划分。

4.1.5 一体化数据资源服务平台方应制定具体操作流程，按流程进行数据采集，保障一体化数据资源服务平台能正确执行数据采集任务。

4.1.6 一体化数据资源服务平台方通过对数据提供方的数据及数据接口进行分析，优化数据采集方案，确定数据采集方式。

4.1.7 数据提供方应遵守一体化数据资源服务平台方制定的管理制度和操作流程。

4.1.8 数据提供方应梳理其公共数据资源目录，明确公共数据更新频率和数据共享属性。

4.1.9 数据提供方应配备操作员，负责数据采集对接工作，按一体化数据资源服务平台数据采集流程采集、上传数据。

4.2 质量要求

数据采集质量应符合 GB/T 36344 的要求，在数据采集周期内，应保证数据的规范性、完整性、准确性、一致性、及时性、可访问性。

a） 规范性。数据符合数据标准、数据模型、业务规划、元数据或权威参考数据的程度。

b） 完整性。相关数据准确和完备的特性，数据项以及数据值存在和缺失的程度。

c） 准确性。数据准确表示其所描述的真实实体（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） 及时性。数据更新符合业务，按照更新周期计算及时率，主要包括天、周、月度、季度、年度。

f） 可访问性。数据能被访问的程度。

4.3 安全要求

4.3.1 数据采集安全保护要求应符合 GB/T 22239 和 GB/T 35274 的规定，对采集数据进行分类分级管理，并根据数据安全措施和技术手段，对数据采集过程进行有针对性地保护，个人信息、敏感数据和重要数

据应加强安全管控措施。
4.3.2 一体化数据资源服务平台方应建立公共数据资源安全管理制度，分级管理、按流程申请、按流程修改。确保数据来源的合法性、正当性、可追溯性，满足公共数据资源采集审计及监测的要求。
4.3.3 一体化数据资源服务平台方应建立数据访问权限管理制度，明确一体化数据资源服务平台方和数据提供方分级权限，安全管理职责分工和工作要求，对数据采集环境、设施和技术采取安全管控措施，数据在整个采集、转化、传输过程中应依据授权使用，不被非法冒充、窃取、算改、抵赖。
4.3.4 一体化数据资源服务平台方应建立公共数据安全监测、安全审计、数字签名、数据加密等技术或手段，对不同数据进行分类并标识，采用安全技术进行安全维护。
4.3.5 一体化数据资源服务平台方应建立应急事件响应机制，编制应急预案，开展应急演练，持续优化应急预案等。
4.3.6 一体化数据资源服务平台方应具备相应的数据服务安全能力，定期对公共数据采集的安全性进行风险评估，并据此制定相应的风险处理计划，及时排查安全漏洞、加固安全技术。
4.3.7 一体化数据资源服务平台方应开展公共数据管理安全培训，编制公共数据安全培训规划，培训公共数据专业知识，开展公共数据安全工作经验交流，宣传公共数据安全知识等。
4.3.8 一体化数据资源服务平台方应明确数据采集过程中个人信息和重要数据的知悉范围和安全管控措施，并采取必要的技术手段和管理措施保证数据不被泄露。
4.3.9 一体化数据资源服务平台方应与数据提供方签订双向保密协议，明确双方数据采集、存储、共享、使用等过程中的职责。
4.3.10 数据提供方应梳理接入应用系统和终端的公共数据，确定接入终端数量、网络带宽和接入地点等信息，制定接入实施方案，明确访问控制、接入认证等安全措施。

5 数据采集方案

5.1 采集方案应包括采集内容、采集频率、采集方式等。
5.2 一体化数据资源服务平台应支持增量更新、全量更新、定时更新、事件触发更新和手动更新等方式。
5.3 一体化数据资源服务平台提供多种标准协议的服务接入方式，包括但不限于数据库抽取、服务网关、消息队列、文件传输、直报系统、标准协议接口。
5.4 公共数据采集实时性要求低的数据可采用定时批量采集的方式，实时性要求高的数据应采用实时采集的方式。
5.5 公共数据量较大、单批量采集可能会造成系统故障的，应使用分批采集。
5.6 宜使用数据提供方系统的备份库作为采集对象，使用备份库时应保证数据一致性和可用性。对于海量数据，宜支持分批或增量读取，宜采用分布式方式对数据源进行读取。

6 数据采集内容

6.1 数据范围

6.1.1政务部门依法履职过程中采集、获取的数据。
6.1.2具有公共职能的企事业单位在提供公共服务和公共管理过程中产生、收集、掌握的各类数据资源，如教育医疗数据、水电煤气数据、交通通信数据等。
6.1.3政府资金资助的专业组织在公共利益领域内收集、获取的具有公共价值的数据，如基础科学研究数据。
6.1.4具有公共管理和服务性质的社会团体掌握的与重大公共利益相关的数据。
6.1.5涉及公共服务领域的其他数据，如社会组织和个人利用公共资源，在提供公共服务过程中收集、产生的涉及公共利益的数据。

6.2 数据类型

6.2.1 按结构化特征、业务归属和产生来源等维度对公共数据进行分类。
6.2.2 结构化数据。对于结构化数据，按业务归属分为：

a） 主数据，用来描述核心业务实体的数据。
b） 基础数据，描述核心业务对象、交易业务的基础信息数据。
c） 事务数据，在业务和流程中产生并记录业务事件的数据。
d） 观测数据，对人、事、物、环境等观测对象，通过观测工具获取的数据。
e） 规则数据，结构化描述业务规则变量的数据。
f） 统计数据，对数据按照统计学方法进行处理加工后，用作业务决策依据的次级数据。

6.2.3 半结构化数据。对于半结构化数据，按产生来源分为 Xml 文档、Json 文档、日志文件、Html 文档、Email 等。
6.2.4 非结构化数据。对于非结构化数据，按产生来源分为文本数据、多媒体数据、空间数据等。

6.3 采集频率

6.3.1 根据数据采集的需要，数据采集频率可分为：数据一次采集、数据实时采集和数据周期采集。
6.3.2 数据一次采集。一次性将所有待采集数据部采集到一体化数据资源服务平台。
6.3.3 数据实时采集。采集响应时间要保证实时性、低延迟，可按秒、分进行数据片采集。
6.3.4 数据周期采集。根据数据产生的业务时间进行数据分片，统一形成周期性的数据片进行数据采集，可按每周、每月、每季、每年等时间周期进行数据片采集。

7 数据映射关系

7.1 数据提供方通过数据库表方式向一体化数据资源服务平台提供数据时，应建立数据提供方数据库表与平台数据库表之间的存储结构映射关系，通过数据库表对接。
7.2 数据提供方以接口方式向一体化数据资源服务平台提供数据时，如返回数据为结构化或半结构化的接口类型，应建立接口返回数据结构与目标数据库之间存储结构映射关系，通过接口进行数据采集。
7.3 数据提供方以接口方式向一体化数据资源服务平台提供数据时，如返回数据为非结构化的接口类型，可将文件通过接口上传后建立文件映射关系表。

8 数据采集方式

8.1 数据采集方式包括终端采集、人工采集、软件系统数据汇聚等。
8.2 终端采集。通过硬件终端、软件终端等方式对物联网传感器数据、互联网数据等进行数据采集。
8.3 人工采集。通过在线填报、离线导入等人工转化方式进行数据采集和导入，如问卷调查、实地调研、资料分析等产生的数据，包括常用的文件交换类型和数据库导出文件。
8.4 软件系统数据汇聚

a） 数据库表交换。以数据库表作为数据资源进行汇聚。通过在数据交换两端部署数据交换组件及交换库，源端数据发生更新后实时通过交换组件推送至源端交换库，由两端交换组件协调双方交换库的同步，目标端通过交换组件从交换库提取数据。
b） 数据接口。以数据接口服务作为数据资源进行汇集，常用的接口方式有 WebService、Restful 等。数据资源提供方调取业务应用系统或数据库中的数据，并封装提供数据接口服务，数据需求方通过数据接口调用获取数据，并把数据采集至前置库中，目标端通过交换组件从前置库提取数据。
c） 文件交换。以电子文件作为数据资源进行汇聚。文件传输可采用 FTP、SFTP 等协议，实现共享

文件数据组装、数据传输、数据解析和数据使用，达到数据交换的目的。

d） 消息队列。以消息发布－订阅方式进行数据汇聚。实现消息的异步发送接收，发布订阅，使得两端的应用解耦合网络传输断点续传，支持分布式消息队列。

9 数据采集过程

9.1 数据采集流程

一体化数据资源服务平台方根据公共数据的采集范围和采集类型，确定数据采集方式，然后进行数据采集准备，实施数据采集，数据质量核查，数据质量提升，原始数据入库，数据更新等，流程如图 1 所示。

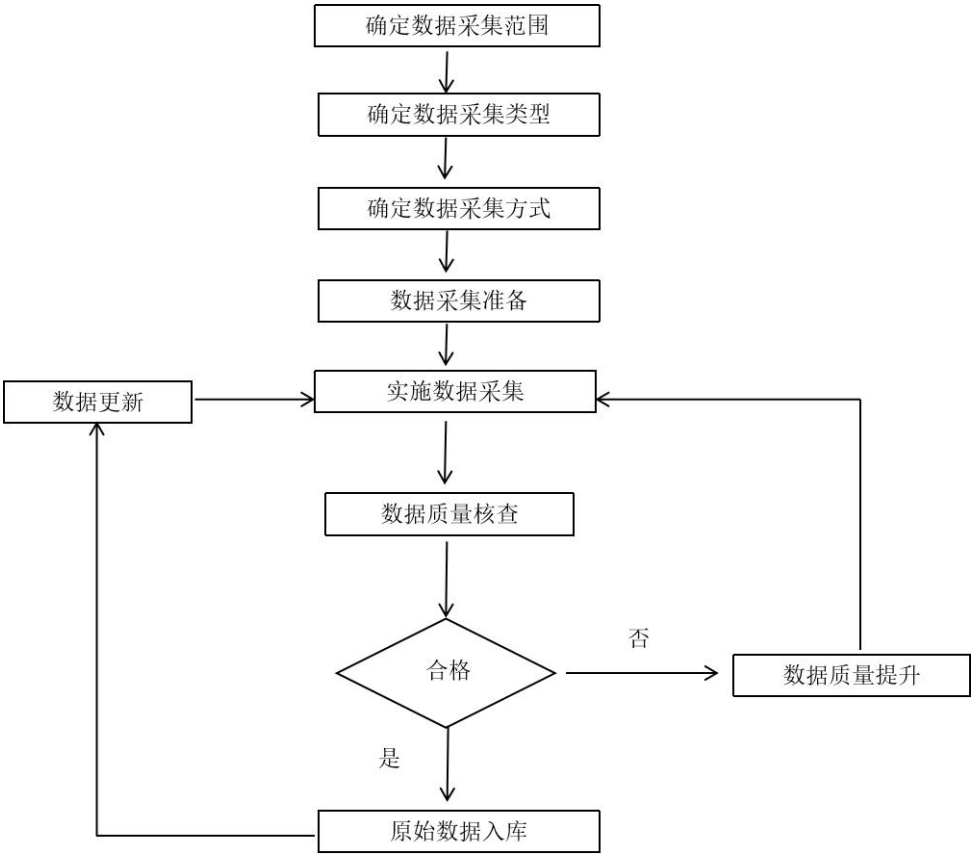

图 1 数据采集流程图

9.2 确定采集方式

一体化数据资源服务平台方根据需要采集数据的范围、类型以及数据的质量和安全要求，综合考虑数据源网络环境、数据采集工具技术路线选型、现有数据采集通道建设情况，从而确定数据采集方式。

9.3 数据采集准备

9.3.1 一体化数据资源服务平台方按照《一体化数据资源体系数据资源目录体系规范》要求，编制公共数据资源目录，并在平台中向数据提供方发布数据采集任务。

9.3.2 数据提供方应分析本单位涉及的公共数据，并将公共数据资源目录规定的数据报送到一体化数据资源服务平台。

9.3.3 数据采集探查主要进行开始实施数据采集前的准备工作，包括以下内容：

a） 数据来源分析：对数据来源业务进行探查分析。

b） 接入方式分析：对原始数据存储位置、提供方式进行分析。
c） 数据结构分析：对数据的含义、类型、长度、结构进行分析。
d） 数据概况分析：对数据内容进行分析，包括数据总数、分布情况、平均值、中位数、最大值、最小值等数据统计分析。
e） 数据关联分析：对数据之间存在的依赖关系、主外键关系进行分析。

9.4 实施数据采集

9.4.1 根据数据采集探查的结果，针对不同的数据源类型，推荐采用以下数据采集方案实施数据采集：
a） 针对结构单一、数据量相对较小的结构化数据，可通过数据库交换、文件交换、数据接口、消息队列等方式进行数据采集。
b） 针对产生的类型丰富、数据量较大的数据，可通过分布式数据接口、分布式流数据收集等方式进行数据采集。
c） 针对海量音视频数据，可通过硬件终端的语音图像识别、编解码等技术转化后进行数据采集。
d） 针对问卷调查、实地调研、资料分析等产生的数据，可通过在线填报、离线导入等人工转化方式进行数据采集。
9.4.2 不应在待采集数据的源系统业务繁忙时实施数据采集，避免数据采集动作影响源系统正常运行。
9.4.3 宜使用源系统的备份库作为采集对象，使用备份库时应保证数据一致性和可用性。
9.4.5 对于数据量较大、单批量采集可能会造成系统故障的，应支持分批或增量读取，并采用分布式方式对数据源进行读取。

9.5 数据质量核查与质量提升

9.5.1 通过对采集的数据总量进行比对，生成数据对比报告，并对采集数据内容进行质量核查，保证获取数据与原始数据数量、数据内容一致。
9.5.2 若经数据质量核查存在数据差异，启动数据质量提升流程，将数据质量评估报告和问题数据清单反馈给数据提供方，待其将异常数据核对修改后重新进行采集。

9.6 原始数据入库

9.6.1 将不进行处理的原始数据采集后存放在公共数据的原始库中。
9.6.2 根据源数据选择合适的数据存储方式对数据进行存储。

9.7 数据更新

9.7.1 数据提供方对原始数据进行更新后，一体化数据资源服务平台方应依照数据采集流程对原始库中更新的数据进行更新采集，并根据数据更新快慢和实时性要求制定不同的采集策略。原始库中更新后的历史数据存放在公共数据中的历史库中。
9.7.2 应支持全量更新和增量更新的数据更新方法：
a） 对存在更新标识的数据应支持增量更新；
b） 对不存在更新标识的数据应支持全量更新。
9.7.3 应支持定时更新、事件触发更新和手动更新的数据更新策略：
a） 对产生呈现周期性规律的数据应支持定时更新策略；
b） 对产生由特定事件触发的数据应支持事件触发更新策略；
c） 对产生无特定规律的数据应支持手动更新策略。
9.7.4 支持实时、定时的数据更新频率，并根据数据变化情况，进行及时和持续更新：
a） 实时产生且实时性要求高的数据应进行实时更新；
b） 实时产生且实时性要求低的数据宜采用定时更新。

10 绩效评价

10.1 一体化数据资源服务平台方应建立公共数据采集的评价机制，对数据采集的效果进行综合分析与评价，并不断改进。
10.2 评价可采用自我评价、用户满意度评价、第三方评价或多方评价相结合等方式进行。
10.3 一体化数据资源服务平台方应通过数据的逐年积累，各绩效指标各年度数据的对比分析，建立标准统一、数据准确、普遍认可的绩效指标体系和评分标准。
10.4 绩效评价指标设计应能反映数据提供方提供数据的基本情况，支持自动生成数据报表模式。

参 考 文 献

[1]《国务院办公厅关于印发全国一体化政务大数据体系建设指南的通知》（国办函〔2022〕102号）

[2]《国务院关于加强数字政府建设的指导意见》（国发〔2022〕14号）

[3]《自治区数字政府改革建设方案》

[4]《自治区数字政府建设三年行动计划（2023—2025年）》

[5]《新疆维吾尔自治区标准化条例》

[6]《新疆维吾尔自治区公共数据管理办法（试行）》