

政务数据汇聚共享规范

Specification for government data aggregation and sharing

(征求意见稿)

2021-xx-xx 发布

2021-xx-xx 实施

北京市市场监督管理局 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 总体架构..... 1

5 汇聚共享数据类别..... 2

 5.1 电子文件..... 2

 5.2 数据库表..... 3

 5.3 数据服务..... 3

6 数据汇聚共享方式..... 3

 6.1 原始数据交换..... 3

 6.2 数据服务调用..... 3

 6.3 数据隐私计算..... 3

7 数据质量要求..... 3

附录 A（规范性） 政务数据汇聚共享业务流程.....5

附录 B（资料性） 数据服务调用步骤.....7

附录 C（资料性） 数据服务调用示例.....8

附录 D（规范性） 数据质量要求..... 12

参考文献..... 12

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。

本文件由北京市经济和信息化局提出并归口。

本文件由北京市经济和信息化局组织实施。

本文件起草单位：

本文件主要起草人：

本文件为首次发布。

政务数据汇聚共享规范

1 范围

本文件规定了政务数据汇聚共享的总体架构、数据类别、共享方式，描述了对数据质量的要求。

本文件适用于各级政务部门的非涉密政务数据通过北京市大数据平台进行汇聚共享的实施和管理。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

政务数据 government data

各级政务部门及其支撑单位在履行职责过程中依法采集、生成、存储、管理的各类数据资源。

[来源：GB/T 38664.1—2020，3.1]

3.2

数据汇聚共享 data aggregation and sharing

各级政务部门因履行职责需要，使用其他政务部门的数据（或服务），以及为其他政务部门提供数据（或服务）的行为。

[来源：GB/T 38664.1—2020，3.2，有修改]

3.3

目录区块链系统 directory blockchain system

利用区块链技术理念，对北京市职责目录、数据目录、库表目录进行统一管控的分布式系统。

3.4

数据（或服务）提供方 data（service）provider

在政务数据汇聚共享过程中，提供数据（或服务）的政务部门。

3.5

数据（或服务）需求方 data（service）demander

在政务数据汇聚共享过程中，申请使用数据（或服务）的政务部门。

3.6

数据目录 data catalog

数据资源和数据项的具体描述，包括数据资源名称、数据资源摘要、数据起始日期、数据更新周期、数据格式、字段名称、数据类型及长度、是否主键、是否非空、数据量等。

4 总体架构

政务数据汇聚共享工作依托北京市大数据平台开展，通过目录区块链系统对政务数据进

行汇聚共享，从提出申请、审批授权、获取共享到评价反馈，各环节操作实时记录上链、全程留痕可溯。总体框架见图 1，详细业务流程按照附录 A 描述执行。

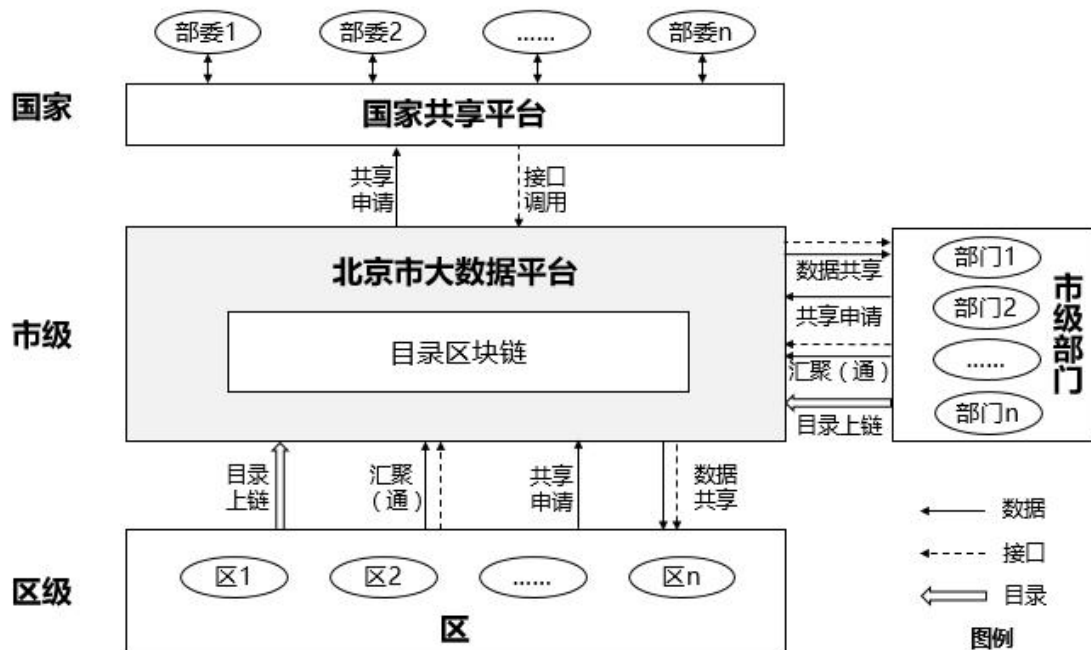


图 1 政务汇聚共享总体架构

5 汇聚共享数据类别

5.1 电子文件

5.1.1 以电子文件作为数据资源进行汇聚共享，常用电子文件的存储格式包括 wps、xml、txt、doc、docx、html、pdf、ppt、csv、xls、xlsx 等。

5.1.2 以电子文件类进行数据汇聚共享时，应遵循以下要求：

- 电子文件名称应规范统一，与目录区块链中数据目录的数据资源名称保持一致；
- 电子文件的存储路径应规范统一，可根据更新的频度和检索效率建立子文件夹，不应随意更改路径；
- 特殊类电子文件应提供必要的说明文档，确保所有文件内容可被正确理解；
- 提供电子文件对账表，明确所汇聚电子文件包含的内容和数量等信息；
- 建立异常反馈机制，通过异常数据反馈表及时解决数据问题。

5.2 数据库表

5.2.1 以数据库表作为数据资源进行汇聚共享，常用数据库存储格式为 oracle、sql server、db2、KingbaseES、access、dbf、dbase、sysbase 等。

5.2.2 以数据库表类进行数据汇聚共享时，应遵循以下要求：

- 数据库表名称应规范统一，一般为“机构简称+数据资源名称缩写”，数据资源名称应与目录区块链中数据目录的数据资源名称保持一致；
- 数据库表字段应与目录区块链中数据目录的数据项保持一致，且应设置主键；
- 数据库表结构应保持稳定，不应随意更改；
- 提供必要的字段说明文档和对应的全部代码表，确保所有数据内容可被正确理解；
- 提供数据对账表，包含数据条数等信息；
- 建立异常反馈机制，通过异常数据反馈表及时解决数据问题。

5.3 数据服务

5.3.1 以数据服务作为数据资源进行汇聚共享，常用的接口方式包括 WebService 和 Restful 等，常用的数据服务格式为 XML 和 JSON 等。

5.3.2 以数据服务类进行数据汇聚共享时，应遵循以下要求：

- a) 应提供详细的服务接口说明文档；
- b) 一个服务接口一般应且只对应一类数据资源；
- c) 服务应是无状态的，两次请求之间无须状态和会话的保持；
- d) 服务地址和参数不应随意变更。

6 数据汇聚共享方式

6.1 原始数据交换

6.1.1 数据库表交换方式

按照数据汇聚共享的场景，支持下列几种模式交换：

- a) 标记位模式：应包含主键、一个标记位字段、推送至库表的时间字段，并且允许共享系统在完成交换之后更改标记位的值。本模式适合大批量、增量模式交换；
- b) 时间戳模式：应包含主键、一个时间戳字段，时间戳日期至少应该精确到毫秒。本模式适合增量模式交换；
- c) 触发器模式：应包含主键、数据表上能建立增删改触发器。本模式适合增量模式交换；
- d) 全量模式：全量模式适合数据库表量少，且每次更新都是全表更新的场景。

6.1.2 电子文件交换方式

提供的电子文件应为常见文件格式，且每次应在固定的文件服务器数据路径下进行文件推送。为满足所有操作系统、文件间系统的通用性，文件资源在完全写入磁盘之前，除非共享系统移动资源，否则其他系统、人员将不能再操作文件。

6.2 数据服务调用

6.2.1 数据接口封装

将各类数据转换为 API 接口，接口设计应遵从统一的接口规范，支持多源异构的数据库格式及协议，包括主流关系型数据库、Hadoop、WebService、FTP、HTTP、自定义协议接口等。

6.2.2 数据服务管理

- 6.2.2.1 为数据提供者开放的数据服务进行相关管理，包括对服务的注册、申请、维护、审核、发布、监控等。
- 6.2.2.2 对于以 API 接口服务方式提供的服务可以在配置下发或获取方式时指定数据范围，系统自动生成接口的授权，访问者只有根据该授权才能获取到数据。
- 6.2.2.3 数据服务调用步骤和服务调用示例详见附录 B 和附录 C。

6.3 数据隐私计算

- 6.3.1 利用多方安全计算、联邦学习、可信执行环境等方式，在数据可用不可见或数据可用不可得的前提下，实现不同来源数据在保持加密状态下进行融合分析运算，并将运算结果进行共享。
- 6.3.2 在密文数据上执行数据操作，避免数据使用方直接接触明文数据，以提高计算过程中的数据安全，确保敏感数据不泄露。

7 数据质量要求

7.1 政务数据汇聚共享时应从数据的可用性、完整性、规范性、一致性和时效性五个方面

保证数据的质量。

7.2 数据质量要求按照附录 D 要求执行。

附 录 A
(规范性)
政务数据汇聚共享业务流程

政务数据汇聚业务流程按照图A.1 执行。

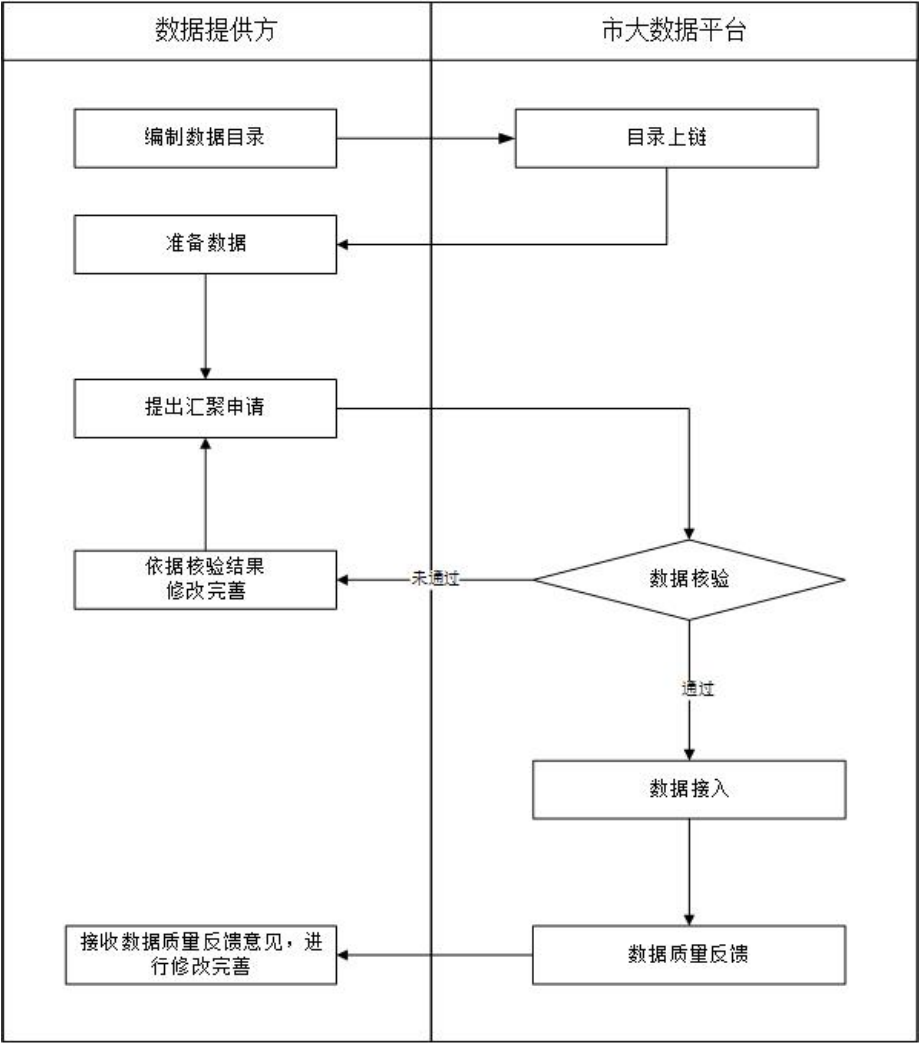


图 A.1 政务数据汇聚业务流程

政务数据共享业务流程按照图 A.2 执行。

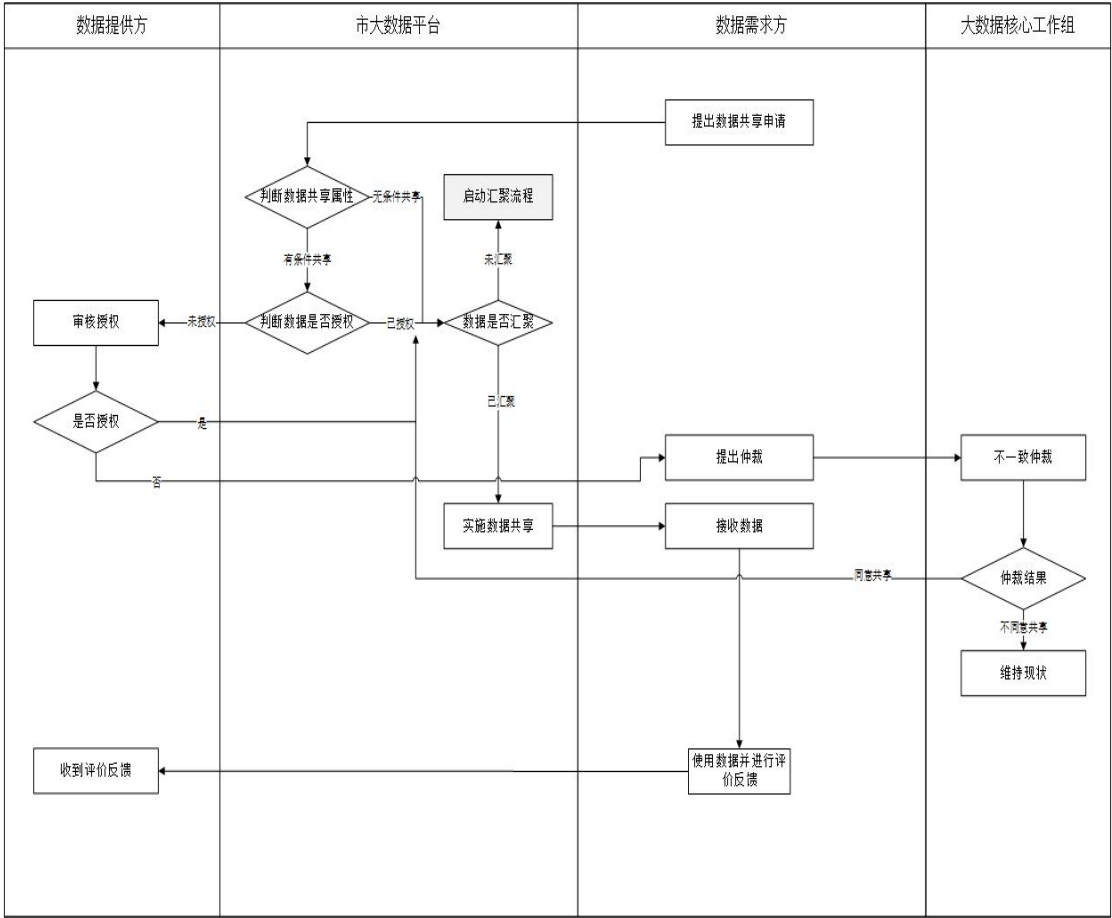


图 A. 2 政务数据共享业务流程

附录 B
(资料性)
数据服务调用步骤

数据服务调用，涉及服务使用方，服务代理方和服务提供方。数据服务调用步骤详见图 B. 1。

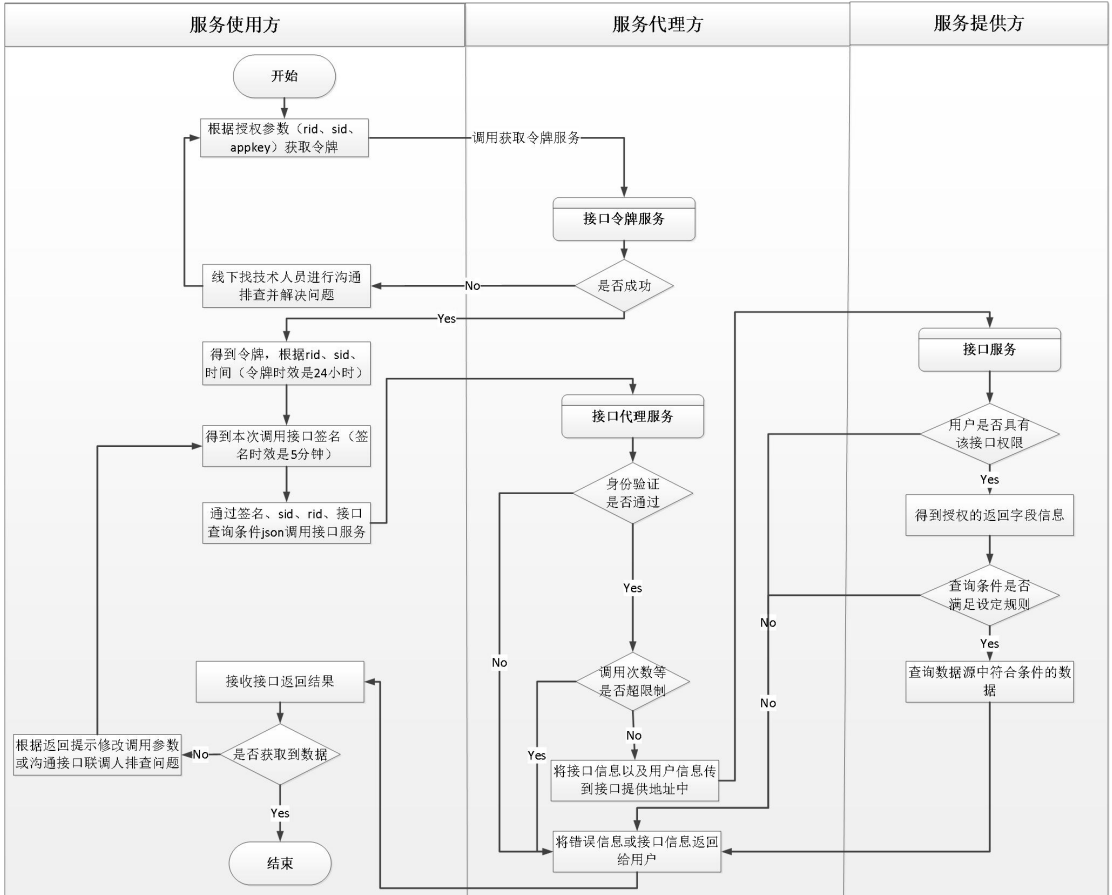


图 B. 1 数据服务调用步骤

B. 1 服务使用方

B. 1. 1 通过授权码，从服务代理端获取令牌。

B. 1. 2 使用令牌，对服务请求者身份标识、服务标识和请求时间进行签名计算，得到签名。

B. 1. 3 将服务请求者身份标识、服务标识、请求时间和签名放入 HTTP 请求的 Header 区，发送请求到服务代理平台。

B. 2 市大数据平台（服务代理方）

获取请求 Header 区的服务请求者身份标识、服务标识、请求时间（_rtime）和签名信息等信息，进行权限验证和调用频率、调用次数、流量的检验。

B. 3 服务提供方

B. 3. 1 获取请求数据进行相关业务处理。

B. 3. 2 根据授权表内容获取输出参数信息过滤输出参数。

B. 3. 3 返回调用结果给服务代理平台（市大数据平台），再由服务代理平台（市大数据平台）返回给服务使用方。

附 录 C

(资料性)

数据服务调用示例

C.1 接口服务入参信息JSON描述

```
{
  "returnType": "json",
  "page": "1",
  "pageSize": "2",
  "whereList":
  [[{'queryField': 'name', 'logical': 'equals', 'queryValue': '测试'}]]
}
```

接口服务入参参数说明见表C.1。

表 C.1 接口服务入参参数说明表

提交方式	POST				
接口协议	HTTP+JSON				
内容类型	application/json				
提交资源数据	名称	是否必须	类型	长度	描述
	returnType	是	string	32	返回数据格式 (json/xml)
	page	否	string	32	页码数 (可以为空, 但要有这个参数), 为空时默认 1
	pageSize	否	string	32	每页最大显示行数 (可以为空, 但要有这个参数), 为空时默认取接口默认值
	whereList	是	Json 数组		查询条件, 下边为 json 对象参数
		名称	类型	长度	描述
		queryField	string	50	查询列
		logical	string	50	查询符 (下表有详细介绍)
		queryValue	string	100	查询值
提交 http header 数据	名称	是否必须	类型	长度	描述
	BJS_sid	是	string	36	服务标识 sid
	BJS_rid	是	string	36	服务 rid
	BJS_sign	是	string	36	签名信息
	BJS_rtime	是	string	36	服务调用时间
返回 Http 状态	200				
返回数据参数 (输出参数)	名称	类型	长度	描述	
	columnsInfo	Json 数组		列信息, 包括类型, 长度, 备注等信息。(下边为 json 对象参数)	
		名称	类型	描述	
		columnName	string	列名称	
		columnComments	string	列备注	
		columnType	string	类型	
		columnLength	string	长度	
	columns	string	1000	列显示字符串	
	page	string	20	页码数	
	maxCount	string	20	每页最大显示行数	
	counts	string	20	查询全部记录数	
	dataList	Json 数组		json 对象中 key 为列名, value 为实际值	

C.2 接口服务出参信息JSON/XML描述

JSON格式

```
{
  "columns": "ZD1, ZD2, ZD3, ZD4, ZD5, ZD6 ",
  "columnsInfo ": [{
    "columnName ": "ZD1",
    "columnComments ": "",
    "columnType ": "DATE",
    "columnLength": "7"
  }, {
    "columnName": "ZD1",
    "columnComments": "字段 1",
    "columnType ": "VARCHAR2 ",
    "columnLength ": "50 "}, {
    "columnName ": "ZD2 ",
    "columnComments ": "字段 2 ",
    "columnType ": "VARCHAR2 ",
    "columnLength ": "18 "}, {
    "columnName ": " ZD3 ",
    "columnComments ": "字段 3",
    "columnType ": "VARCHAR2 ",
    "columnLength ": "10 "}, {
    "columnName ": "ZD4",
    "columnComments ": "字段 4",
    "columnType ": "DATE ",
    "columnLength ": "7 "}, {
    "columnName ": "ZD5 ",
    "columnComments ": "字段 5 ",
    "columnType ": "DATE ",
    "columnLength": "7"}, {
    "columnName": "ZD6",
    "columnComments": "字段 6",
    "columnType ": "VARCHAR2 ",
    "columnLength ": "6 "}],
  "dataList": [
    {"ZD1": "XXX", "ZD2": "1111111", "ZD3": "XX", "ZD4 ": "2020 - 01 - 01 ", "ZD5 ": "1998 - 01 - 01 ", "ZD6 ": ""},
    {"ZD1": "XXX", "ZD2": "1111111", "ZD3": "XX", "ZD4 ": "2020 - 01 - 01 ", "ZD5 ": "1998 - 01 - 01 ", "ZD6 ": ""}
  ],
  "counts ": "1000 ",
  "page ": 1,
  "maxCount ": 2}

```

XML格式

```
<?xml version="1.0" encoding="UTF-8"?>
<columns> ZD1, ZD2, ZD3, ZD4, ZD5, ZD6</columns>
<columnsInfo>
<columnInfo>
<columnName>ZD1</columnName>
<columnComments></columnComments>

```

```

<columnType>DATE</columnType>
<columnLength>7</columnLength>
</columnInfo>
<columnInfo>
<columnName>ZD1</columnName>
<columnComments>字段 1</columnComments>
<columnType>VARCHAR2</columnType>
<columnLength>50</columnLength>
</columnInfo>
<columnInfo>
<columnName>ZD2</columnName>
<columnComments>字段 2</columnComments>
<columnType>VARCHAR2</columnType>
<columnLength>18</columnLength>
</columnInfo>
<columnInfo>
<columnName>ZD3</columnName>
<columnComments>字段 3</columnComments>
<columnType>VARCHAR2</columnType>
<columnLength>10</columnLength>
</columnInfo>
<columnInfo>
<columnName>ZD4</columnName>
<columnComments>字段 4</columnComments>
<columnType>DATE</columnType>
<columnLength>7</columnLength>
</columnInfo>
<columnInfo>
<columnName>ZD5</columnName>
<columnComments>字段 5</columnComments>
<columnType>DATE</columnType>
<columnLength>7</columnLength>
</columnInfo>
<columnInfo>
<columnName>ZD6</columnName>
<columnComments>字段 6</columnComments>
<columnType>VARCHAR2</columnType>
<columnLength>6</columnLength>
</columnInfo>
</columnsInfo>
<dataList>
<data>
<ZD1>XXX</ZD1>
<ZD2>111111</ZD2>
<ZD3>XXX</ZD3>
<ZD4>2020-01-01</ZD4>
<ZD5>1998-01-01</ZD5>
<ZD6>XXX</ZD6>
</data>

```

```
</dataList>  
<counts>10000</counts>  
<page>1</page>  
<maxCount>2</maxCount>
```

附 录 D

（规范性）

数据质量要求

D.1 数据可用性

D.1.1 政务数据汇聚共享应确保数据可读、可理解、可用。

D.1.2 有信息系统支撑的数据应提供结构化文件，并在汇聚数据时同步提供数据字典和码表，确保数据的可读可理解。

D.1.3 所提供的数据应保持独立可用，避免多项业务数据混合提供。

D.1.4 通过接口方式对接的，数据提供方要遵循接口传输规范，具有完整的日志记录，保证数据安全可用。

D.2 数据完整性

D.2.1 目录完整

数据汇聚共享前应确认形成完整的数据目录，按照实际汇聚数据对北京市目录区块链系统中数据目录进行完善。

D.2.2 字段完整

应涵盖该数据的所有有效字段。

示例：“户籍人口登记信息”包含“姓名、出生日期、性别、身份证号码、籍贯、家庭住址、曾用名”字段，汇聚数据应包含全部字段。

D.2.3 释义完整

应确保字段取值所配套的字典表、码表的完整性。

示例：“SEX”字段对应的字典名称“性别”，数据内容为“0”、“1”，“0”对应的实际内容为“男”，“1”对应的实际内容为“女”，汇聚数据应包含所有字段的完整解释。

D.2.4 周期完整

应覆盖该数据自采集日期起至今的全量历史数据，同时数据字段中应包含数据入库时间。

示例：“户籍人口登记信息”采集起始时间为2006年，则应汇聚自2006年至今的全量历史数据。

D.3 数据规范性

D.3.1 格式规范

D.3.1.1 有信息系统支撑的应提供结构化数据，以结构化或半结构化文件方式提供的（如xls、xlsx、csv等），其中xml、json可以为英文标签，但要提供数据项对应中文说明；每个csv文件第一行都为数据项名称，每类数据csv文件结构必须保持不变。

D.3.1.2 通过接口方式对接的，数据提供方要遵循接口传输规范，具有完整的日志记录，保证数据安全可用。

D.3.2 内容规范

D.3.2.1 数据的核心（非空）字段不能为空值。

示例：“户籍人口登记信息”中的“姓名”字段不能为空。

D.3.2.2 不应包含因业务或技术原因产生的冗余数据。

示例：同类数据中不应包含2条完全相同的记录。

D.3.2.3 不应包含错误数据。错误数据包括以下情况：

a) 无效测试数据。系统建设或测试过程中残留的、无实际业务意义的测试数据。

示例：数据中包含多条“test”、“111111”等无效记录。

b) 非法格式数据。字符类型、长度等不满足格式规范约束的数据。

示例：“年龄”为“%%”；“身份证号码”位数为19位。

c) 非法值数据。数据中包括多余的空格、乱码、全角、繁体等错误数据。

示例：“姓名”为“张三”。；“电话号码”为全角数字例如“1 3 1 2 3 4 5 6 7 8 9”。

d) 范围溢出数据。超出字段取值范围的数据。

示例：“年龄”为“300”。

e) 逻辑错误数据。明显不符合合理业务逻辑的数据。

示例：“身份证号码”7到10位（表示出生年份）为“1790”。

D.4 数据一致性

D.4.1 数据目录一致性

数据汇聚共享以各部门目录区块链中数据目录为基准，应确保数据资源名称、数据项名称、数据格式等与“上链”数据目录完全一致，以非结构化文件方式提供的（如PDF、JPG、Word等），需在目录区块链“数据项”中填写为“附件”。大数据平台获取到数据后会与“上链”数据目录进行核验对比，核验通过才算完成汇聚。

D.4.2 数据内容一致性

D.4.2.1 不应存在数据名称相同、数据项（字段）不同的情况。

示例：两类数据名称均为“北京市平均薪资水平”，其中一类数据的字段为“教育行业平均薪资、公安行业平均薪资等”，另一类数据的字段为“年份、薪资水平、详情”。

D.4.2.2 不应存在数据项（字段）相同、数据名称不同的情况。

示例：三类数据的字段均为“姓名、身份证号码、性别、年龄、居住地”，但数据名称分别为“个人基本信息”、“个人信息”、“个人基本数据”。

D.4.2.3 不应存在数据名称和数据项（字段）相同、数据内容不同。

示例：来自两个信息系统的“2007年平均薪资”数据，分别为“2000”和“2500”。

D.5 数据时效性

D.5.1 更新周期

市大数据平台汇聚的数据应与各政务部门生产系统数据的更新周期保持同步更新。

D.5.2 更新方式

D.5.2.1 增量方式：一般情况下以增量方式对汇聚数据进行更新。

D.5.2.2 全量方式：如有特殊原因需要全量更新的，根据具体情况判断。但数据量较大的数据不建议以全量方式更新，例如数据量条数在100万量级以上的。

参 考 文 献

- [1] GB/T 7027-2002 信息分类和编码的基本原则与方法
 - [2] GB/T 21062.1-2007 政务信息资源交换体系 第1部分：总体框架
 - [3] GB/T 21062.2-2007 政务信息资源交换体系 第2部分：技术要求
 - [4] GB/T 21062.3-2007 政务信息资源交换体系 第3部分：数据接口规范
 - [5] GB/T 21062.4-2007 政务信息资源交换体系 第4部分：技术管理要求
 - [6] GB/T 38664.2-2020 信息技术 大数据 政务数据开放共享 第2部分：基本要求
-