

SZSD

数 字 山 东 工 程 标 准

SZSD03 0002—2024

农业农村大数据 数据采集标准规范

Agriculture and rural big data—Data acquisition specifications

2024 - 04 - 15 发布

2024 - 06 - 01 实施

淄博市大数据局 发 布

目 次

前 言 II

1 范围 3

2 规范性引用文件 3

3 术语和定义 3

 3.1 结构化数据..... 3

 3.2 半结构化数据..... 3

 3.3 非结构化数据..... 3

 3.4 关系型数据库..... 4

 3.5 数据源..... 4

 3.6 数据湖..... 4

 3.7 数据采集..... 4

 3.8 原始库..... 4

 3.9 标准库..... 4

 3.10 数据探查..... 4

4 缩略语 4

5 数据采集总体框架 5

 5.1 采集内容..... 5

 5.1.1 数据范围..... 5

 5.1.2 数据类型..... 5

 5.2 采集流程..... 5

6 采集技术规范 6

 6.1 采集技术要求 6

 6.2 采集方式 7

 6.2.1 人工采集 7

 6.2.2 系统采集 7

 6.3 采集配置管理..... 7

 6.4 采集过程管理..... 7

7 数据采集质量控制 8

 7.1 数据质量控制原则..... 8

 7.2 数据质量控制方式..... 8

 7.3 数据质量评价方法..... 8

8 数据采集安全控制 8

参 考 文 献 9

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。

本文件由淄博市农业农村局提出，淄博市大数据局归口。

本文件起草单位：淄博市数字农业农村发展中心、中国农业科学院农业信息研究所。

本文件主要起草人：杜志钢、崔运鹏、刘海明、刘娟、徐学亮、陈丽、王坤、鲜建国、车呈瑾、张亚男、王晓瑞、王浩东、魏春英、王末、王婷、李欢。

农业农村大数据 数据采集标准规范

1 范围

本文件规定了淄博市农业农村数据采集的总体要求与管理要求，从采集内容、采集过程及技术进行规范。

本文件适用于淄博市农业农村智慧大脑及各农业农村下属单位开展数据采集工作。

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 35295—2017 信息技术 大数据 术语

GB/T 36344—2018 信息技术数据质量评价指标

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

3 术语和定义

GB/T 35295—2017界定的以及下列术语和定义适用于本文件，为了方便使用，以下重复列出了GB/T 35295—2017中的某些术语和定义。

3.1

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。

注：结构化数据也称作行数据，是由二维表结构来逻辑表达和实现的数据，严格地遵循数据格式与长度规范，主要通过关系型数据库进行存储和管理。

[来源：GB/T 35295—2017，定义2.2.13]

3.2

半结构化数据 semi-structured data

自描述的数据，是介于结构化和非结构化之间的数据。

注：半结构化数据具有一定的结构性，如：JSON文件，XML文件等。

3.3

非结构化数据 unstructured data

不具有预定义模型或未以预定义方式组织的数据。

注：非结构化数据的数据结构不规则或不完整，不方便用数据库二维逻辑表来表现，包括所有格式的办公文档、文本、图片、各类报表、图像和音频/视频信息等。

[来源：GB/T 35295—2017 定义2.1.25]

3.4

关系型数据库 relational database

即关系型数据库管理系统，是利用数据概念实现对数据处理的算法，达到对数据及其快速的增、删、改、查操作。关系型数据库模型是把复杂的数据结构归结为简单的二元关系（即二维表格形式）。

3.5

数据源 data source

提供一种所需要数据的器件或原始媒体，在数据源中存储了所有建立数据库连接的信息。

3.6

数据湖 data lake

数据湖是一类存储数据自然/原始格式的系统或存储，通常是对象块或者文件。数据湖中包括来自于关系型数据库中的结构化数据（行和列）、半结构化数据（如CSV、日志、XML、JSON）、非结构化数据（如email、文档、PDF等）和二进制数据（如图像、音频、视频）。

3.7

数据采集 data acquisition

从数据源中得到原始数据，将各种类型原始数据统一接入，汇聚存储至数据湖中的过程。

3.8

原始库 operational data store

存储由源系统采集而来、未经处理的原始数据。

3.9

标准库 standard data store

存储由原始数据经过基础清洗、标准化数据。

3.10

数据探查 data profiling

对数据的结构、内容、关系和来源进行调研、探查，以达到确定数据采集通道畅通、认识数据的目的。

4 缩略语

下列缩略语适用于本文件。

API：应用程序接口（Application Programming Interface）

JSON：一种轻量级的数据交换格式(JavaScript Object Notation)

XML：可扩展标记语言(Extensible Markup Language)

HTTP：超文本传输协议（Hypertext Transfer Protocol）

VPN：虚拟专用网络（Virtual Native Network）

5 数据采集总体框架

5.1 采集内容

5.1.1 数据范围

农业农村大数据采集范围包括但不限于基础数据、专题数据、业务数据和其他数据四大类。

- a) 基础数据：包括农业自然资源、环境、气象、地理等基础数据；
- b) 专题数据：针对特定主题或领域的的数据，如农产品市场专题数据、乡村治理专题数据等；
- c) 业务数据：与农业农村部门相关的各种业务数据，如农业生产管理、农业技术推广等数据；
- d) 其他数据：其他农业农村相关的数据，如农村人口变化、农民收入水平、农村社会保障、农村人口结构、农村人口教育程度等数据。

5.1.2 数据类型

考虑到农业农村大数据的多源性和异构性等特点,采用混合分类法进行分类。首先,采用线分类法,按照结构化特征进行大类划分；其次,针对某一大类数据,采用面分类法,按照业务归属、数据来源、数据时限等进行小类划分。

- a) 结构化数据：对于结构化数据，根据业务归属，可以分为种植业、畜牧业、农技推广等数据等；
- b) 半结构化数据：对于半结构化数据，按照产生来源，可分为 XML 文档、JSON 文档、邮件、日志文件等数据；
- c) 非结构化数据：对于非结构化数据，按照产生来源可分为文本数据、多媒体数据、空间数据等。

5.2 采集流程

根据数据的采集范围和采集类型，确定数据采集方式，进行模拟探查，实施数据采集、质量核查及异常处理，完成原始数据入库，经过数据规整后存入标准数据库。总体框架如图1所示。

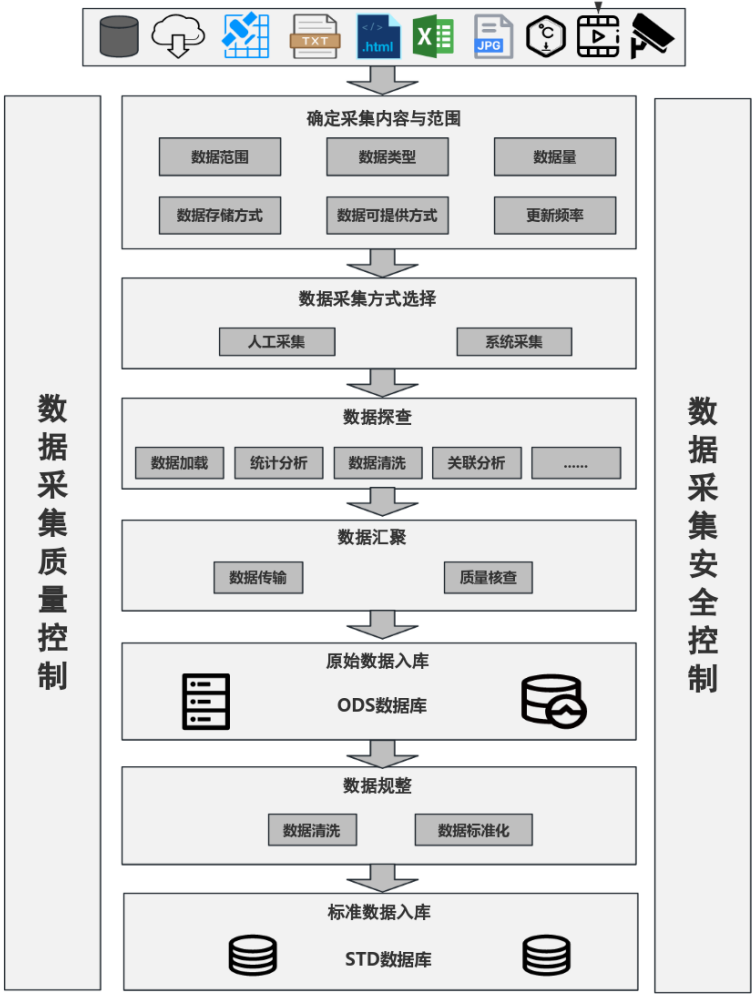


图1 数据采集总体框架图

数据采集总体框架包括：

- a) 确定数据采集内容与范围：明确需要采集的数据范围、数据类型、数据量、更新方式等；
- b) 选择数据采集方式：根据数据范围、类型、数据量等因素，综合安全和质量要求，确定数据接入方式；
- c) 数据探查：快速、直观了解数据质量现状，为数据采集提供基础；
- d) 实施数据采集：数据采集存入原始数据库，应根据数据源特征选择合适的数据存储方式；
- e) 数据规整：对原始数据进行必要地转换、清洗、填充、平滑等操作后，存入标准数据库；
- f) 采集数据质量控制，数据质量控制贯穿整个采集过程，主要质量控制方式包括：数据清洗、数据验证、数据标准化、数据校准等；
- g) 采集数据安全控制，数据安全控制也贯穿整个采集过程，从数据源的选择、采集方式的设计、数据传输和存储，到数据的使用和共享，都需要进行严格地安全控制。

6 采集技术规范

6.1 采集技术要求

数据采集技术应具备复杂网络环境下，不同异构数据源之间的高速、稳定、弹性收缩的数据移动及同步能力。

- a) 针对结构简单、数据量较小的结构化数据，可通过数据库表、文件、Web 服务（WebService）、HTTP、消息订阅或发布等技术进行数据采集；
- b) 针对传感器、智能手机、网络等渠道产生的类型丰富、数据量较大的数据、可通过分布式系统接口，分布式流数据收集、网络爬虫等技术进行数据采集；
- c) 针对摄像头等音视频设备产生的海量数据，可通过编码压缩后进行数据采集；
- d) 针对实地调研、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。

6.2 采集方式

6.2.1 人工采集

数据采集提供人工录入、人工导入数据界面，可对批量结构化数据进行系统导入，录入内容支持表单和表格类数据自定义设计。录入时用户可通过界面直接录入和编辑，也可以通过excel导入的方式批量上传数据，用户录入数据后支持数据汇总和存储。

6.2.2 系统采集

- a) 终端采集：通过硬件终端、软件终端、网络爬虫等方式对物联网传感器数据、互联网数据等进行采集；
- b) 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同步，目标端通过交换组件从交换库提取数据；
- c) 数据接口：以数据接口服务作为数据资源进行汇聚，常用的接口方式有 Webservice、Restful，并以 XML、JSON 等格式进行服务。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；
- d) 文件交换：以电子文件作为数据资源进行汇聚，常用的电子文件类型有 wps、xml、txt、doc、docx、html、csv、xls、xlsx 等。通过前置机的共享目录或 FTP 服务，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；
- e) 消息队列：以消息发布-订阅方式进行数据汇聚，可实现消息的异步发送接收，发布订阅，使得两端的应用解耦（减少或解除应用程序之间的耦合度）和网络传输断点续传，支持分布式消息队列。

6.3 采集配置管理

提供对多种数据接入方式进行配置的管理功能，用户可以通过配置管理配置数据接入方式、数据来源、数据范围、数据资源获取频率、数据存储位置、数据存储策略等采集要素信息，可按配置好的方式实现数据自动采集。

6.4 采集过程管理

为满足对采集整体过程可视、可控、可测等全流程管理要求，满足低技术要求和低成本的运维和采集实施要求，数据采集需提供全流程可视化管理和操作界面，为用户提供数据源管理、围绕采集任务或流程的构建、启停、过程监控、异常通知和恢复管理、采集结果查看等功能或界面。

7 数据采集质量控制

7.1 数据质量控制原则

数据质量控制贯穿整个采集过程，需遵循但不限于以下原则：

- a) 数据完整性，应包含数据规则要求的数据的必要元素：数据格式、数据指标、报送频率、报送时间等；
- b) 数据准确性，应真实反映数据所描述的实体；
- c) 数据时效性，应保证数据发生变化后及时更新；
- d) 数据一致性，应保证数据与其特定上下文中使用的数据无矛盾；
- e) 数据可访问性，应保证数据在需要时被安全访问；
- f) 数据可追溯性，应明确数据提供方及提供方式，保证数据能够被追踪和管理。

7.2 数据质量控制方式

数据质量控制方式包括但不限于以下方式：

- a) 数据验证：数据质量控制的首要步骤，确保从源系统采集的数据是有效和合法的。包括格式验证、范围验证和类型验证等，如验证手机号码是否符合特定格式、验证日期是否在合理范围内；
- b) 数据清洗：识别和纠正数据中的错误、拼写错误、不一致性和不完整性。包括自动或手动修复拼写错误、填充缺失数据、处理异常值等；
- c) 数据去重：检测并移除数据采集过程中出现的重复数据记录，确保数据的唯一性；
- d) 数据标准化：将数据转换为统一的格式和单位，以便于比较和分析。例如，将所有日期数据格式化为相同的日期格式；
- e) 数据完整性检查：检测数据中的缺失或不完整的部分。可以通过验证必需字段是否包含数据，或者通过检查数据记录中的关键字段是否为空来实现；
- f) 数据审计：追踪和记录数据的变化和访问历史。通过数据审计，可以确定数据何时被创建、修改或删除，以便追溯数据的来源和变更历史。

7.3 数据质量评价方法

数据质量评价方法可分为定性评价法和定量评价法：

- a) 定性评价法主要通过人工的方式对数据进行审查和分析，判断数据的可靠性和准确性；
- b) 定量评价法主要通过一系列的指标和算法对数据进行量化评估，从而得出数据质量的评分。

8 数据采集安全控制

数据采集安全贯穿于整个数据采集整个过程中，应符合主管部门的安全要求，包括但不限于：

- a) 应符合 GB/T 35273 和 GB/T 22239 的规定，对采集的数据进行分级分类，针对个人信息、敏感信息和重要信息，采取更加严格的安全控制措施；
- b) 数据在整个采集、清洗、转化、传输过程中应防止被非法冒充、窃取、篡改；
- c) 对于不同类型数据，应进行分类标识，按照不同的安全等级进行安全保护；
- d) 对于不同安全等级的数据，应实施权限控制；
- e) 采用密码、VPN、数字签名等技术维护数据安全；
- f) 制定数据采集操作规程，规范数据采集的数据格式、方法和流程等；
- g) 建立安全管理规范，避免人为因素造成数据泄露、损坏等安全事故。

参 考 文 献

- [1]DB37/T 2874-2016 农业物联网平台基础数据采集规范
 - [2]DB37/T 3432-2018 农业大数据 数据处理基本要求
 - [3]GB/T 22239-2019 信息安全技术—网络安全等级保护基本要求
 - [4]GB/T 35273 -2018 信息安全技术—个人信息安全规范
 - [5]GB/T 35873-2018 农产品市场信息采集与质量控制规范
 - [6]农业综合标准化工作指南
-