|  |  |
| --- | --- |
| ICS  | 07.040 |
| CCS  | A 75 |

|  |
| --- |
|  SZSD |

数字山东工程标准

SZSD XXXX—2021

威海市统一地址库 地名地址数据采集入库规范

Weihai unified address database—Specification for geographical name and address collection and storage

XXXX - XX - XX发布

XXXX - XX - XX实施

威海市大数据局  发布

目次

[前言 II](#_Toc90970567)

[1 范围 1](#_Toc90970568)

[2 规范性引用文件 1](#_Toc90970569)

[3 术语和定义 1](#_Toc90970570)

[4 数据属性描述 2](#_Toc90970571)

[4.1 概述 2](#_Toc90970572)

[4.2 数据项名称 2](#_Toc90970573)

[4.3 英文名称 2](#_Toc90970574)

[4.4 数据项定义 2](#_Toc90970575)

[4.5 数据类型 2](#_Toc90970576)

[4.6 数据格式 2](#_Toc90970577)

[4.7 约束 3](#_Toc90970578)

[4.8 备注 3](#_Toc90970579)

[5 地址库建库 3](#_Toc90970580)

[6 数据采集 3](#_Toc90970581)

[6.1 采集原则 3](#_Toc90970582)

[6.2 采集范围 4](#_Toc90970583)

[6.3 采集内容 4](#_Toc90970584)

[7 数据处理 6](#_Toc90970585)

[7.1 概述 6](#_Toc90970586)

[7.2 数据清洗 6](#_Toc90970587)

[7.3 数据比对 6](#_Toc90970588)

[7.4 数据标准化 7](#_Toc90970589)

[8 数据入库 7](#_Toc90970590)

[8.1 入库流程 7](#_Toc90970591)

[8.2 入库前检查 7](#_Toc90970592)

[8.3 数据导入 7](#_Toc90970593)

[8.4 入库后检查 7](#_Toc90970594)

[9 数据更新 8](#_Toc90970595)

[9.1 更新方式 8](#_Toc90970596)

[9.2 更新要求 8](#_Toc90970597)

[10 数据安全 8](#_Toc90970598)

[参考文献 9](#_Toc90970599)

1. 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。

本文件由威海市大数据局提出、归口并组织实施。

本文件起草单位：威海市大数据中心、威海市委政法委、威海市自然资源和规划局、威海市公安局、威海市民政局、威海市住房和城乡建设局、山东新一代标准化研究院有限公司、威海微网信息技术服务有限公司。

本文件主要起草人：朱崇东、张剑嵩、段升莹、张广琳、连慧超、林栋、于国华、段志伟、刘军涛、赵晓华、蒋鹏、毕明波、李杨、史丛丛、王溪。

威海市统一地址库 地名地址数据采集入库规范

* 1. 范围

本文件规定了统一地址库中地名地址数据的数据属性描述、地址库建库、数据采集、数据处理、数据入库、数据更新和数据安全等要求。

本文件适用于统一地址库中地名地址数据的采集、建库、入库和更新。

* 1. 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 2260 中华人民共和国行政区划代码

GB/T 2312 信息交换用汉字编码字符集 基本集

GB/T 7408 数据元和交换格式 信息交换 日期和时间表示法

GB/T 19488.1—2004 电子政务数据元 第1部分：设计和管理规范

GB/T 19488.2—2008 电子政务数据元 第2部分：公共数据元目录

GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求

GB/T 35273—2020 信息安全技术 个人信息安全规范

DB37/T 3523.2—2019 公共数据开放 第2部分：数据脱敏指南

* 1. 术语和定义

下列术语和定义适用于本文件。

地名地址 address

由标识和定位所处位置的一系列地名组合而成的结构化地址信息。

[来源：GB/T 39609，3.1]

门址 door number

表示院落、独立门户编号的文字描述。

楼址 building number

表示单幢建筑编号的文字描述。

楼层 floor

建筑物的分层。

户室 room

建筑物中的一套或一个房间。

* 1. 数据属性描述
		1. 概述

地名地址数据集的属性描述统一规定采用6项属性，见表1。

1. 地名地址数据属性

| 序号 | 数据元属性名称 | 约束 |
| --- | --- | --- |
| 1 | 数据项名称 | 必选 |
| 2 | 英文名称 | 必选 |
| 3 | 数据项定义 | 必选 |
| 4 | 数据类型及格式 | 必选 |
| 5 | 约束 | 必选 |
| 6 | 备注 | 可选 |

* + 1. 数据项名称

数据项名称是赋予数据项属性的标记，名称应唯一，以字符串形式表示。

* + 1. 英文名称

英文名称是数据项目属性的标识，名称应唯一，采用英文字母大写的形式表示。

* + 1. 数据项定义

数据项定义是描述一个数据项的本质特性并使其区别于所有其他数据项的陈述。

* + 1. 数据类型

数据类型描述要求应符合GB/T 19488.1—2004中5.5.2，见表2。

1. 数据类型

| 数据类型 | 说明 |
| --- | --- |
| 字符型 | 通过字符形式表达的值的类型 |
| 数字型 | 通过从“0”到“9”数字形式表达的值的类型 |
| 日期型 | 通过YYYYMMDD或YYYY-MM-DD的形式表达的值的类型，应符合GB/T 7408 |
| 日期时间型 | 通过YYYYMMDD hhmmss或YYYY-MM-DD hh:mm:ss的形式表达的值的类型，应符合GB/T 7408 |
| 布尔型 | 两个且只有两个表明条件的值，如On/Off、True/False |
| 二进制 | 上述无法表示的其他数据类型，如图像、音频、视频、文件等 |

* + 1. 数据格式

数据格式描述要求符合GB/T 19488.1—2004中5.5.3，见表3。

1. 数据格式

|  |  |
| --- | --- |
| 符号 | 符号说明 |
| a | 字母字符 |
| n | 数字字符 |
| an | 字母数字字符 |
| m(m为自然数) | 定长m个字符（字符集默认为GB/T 2312） |
| ..ul | 长度不确定的文本 |
| ..p,q（p,q均为自然数） | 最长p个数字字符，小数点后q位 |
| .. | 从最小长度到最大长度，前面附加最小长度，后面附加最大长度 |
| YYYYMMDD hhmmss或YYYY-MM-DD hh:mm:ss | “YYYY”表示年份，“MM”表示月份，“DD”表示日期，“hh”表示小时，“mm”表示分钟，“ss”表示秒，可视实际情况组合使用 |
| 1. an5(aannn) 表示定长5个字母数字字符，前2个为字母字符，后3个为数字字符
2. n..17，2 表示最长17个字符（不含小数点），其中小数点后2位
3. an3..8 表示最大长度为8，最小长度为3的不定长字母数字字符
 |

* + 1. 约束

数据项的约束条件，必填或选填。

* + 1. 备注

数据项的附加注释说明。

* 1. 地址库建库

统一地址库建库包括但不限于以下要求：

1. 概念设计：
	1. 应对归集地名地址数据进行归类、分析等；
	2. 应建立不依赖于数据库软硬件环境的数据概念模型。
2. 功能设计：统一地址库的功能设计，应依据各部门的数据应用需求情况，满足相关服务要求；
3. 逻辑设计：
	1. 应建立统一地址库逻辑模型，明确数据组织形式；
	2. 应建立不同数据间的逻辑关联关系。
4. 物理设计：应依据统一地址库逻辑模型，进行统一地址库的物理设计，包括但不限于存储结构、存储路径、系统配置等内容；
5. 安全设计：安全设计应符合第10章的相关规定。
	1. 数据采集
		1. 采集原则

数据采集原则主要包括：

1. 完整性：地名地址内容和属性值应完整，无遗漏、多余或重复的现象；
2. 可追溯性：地名地址应记录可追溯的变化信息；
3. 唯一性：地名地址应具有唯一、固定的标识。
	* 1. 采集范围

统一地址库包括历史的、现势的地名地址数据，采集的范围包括但不限于：

1. 自然地理实体名称：山、河、湖、海、岛礁、沙滩、岬角、海湾、水道、地形区等名称；
2. 行政区划名称：各级行政区域和各级人民政府派出机构所辖区域名称；
3. 居民地名称：城镇、区片、开发区、自然村、片村、农林牧渔点及街、巷、居民区、楼群（含楼、门号码）、建筑物、单元、楼层、户室等名称；
4. 各专业部门使用的特殊地名名称：具有地名意义的台、站、港、场、路等名称以及名胜古迹、纪念地、游览地、企业事业单位等名称。
	* 1. 采集内容

地名地址信息数据集见表4。

1. 地名地址信息数据集

| 序号 | 数据项名称 | 英文名称 | 数据项定义 | 数据类型及格式 | 约束 | 备注 |
| --- | --- | --- | --- | --- | --- | --- |
|  | 省 | PROVINCE | 所在省的名称 | an..50 | 必填 |  |
|  | 市 | CITY | 所在地级市的名称 | an..50 | 必填 |  |
|  | 区县 | COUNTY | 所在县、县级市、区的名称 | an..50 | 必填 |  |
|  | 行政区划代码 | XZQHCODE | 设有国家政权机关的各级地区的数字代码。 | n6 | 必填 | GB/T 2260中的行政区划数字代码 |
|  | 乡镇（街道） | TOWN | 所在的街道、乡镇的名称 | an..100 | 必填 | GB/T 19488.2中的“乡镇（街道）名称” |
|  | 乡镇（街道）代码 | XZCODE | 乡镇街道的代码 | n3 | 必填 | GB/T 39609中的“乡镇（街道）代码” |
|  | 社区 | COMMUNITY | 所在社区、居委会、行政村的名称 | an..100 | 必填 |  |
|  | 院落 | VILLAGE | 所在居民小区、自然村、厂区、学校的名称 | an..100 | 选填 |  |
|  | 院落门址 | VILLAGEADD | 院落的门牌地址 | an..20 | 选填 |  |
|  | 街路巷 | STREET | 所在街、路、巷等的名称 | an..100 | 选填 | GB/T 19488.2中的“街路巷” |
|  | 街巷院落代码 | JXYLCODE | 街巷院落的代码 | n8 | 必填 | GB/T 39609中的“街巷院落代码” |
|  | 建筑门楼 | DOOR | 所在门址、楼址、建筑物的名称 | an..100 | 必填 |  |
|  | 建筑门楼代码 | JZMLCODE | 建筑门楼代码 | n3 | 必填 | GB/T 39609中的“建筑门楼代码” |
|  | 单元 | UNIT | 所在单元的编号 | an..10 | 选填 | 汉字、字母、数字或其组合，例如一单元、二单元等 |
|  | 单元代码 | UNITCODE | 单元的顺序编码 | n2 | 选填 | GB/T 39609中的“单元代码” |
|  | 楼层 | FLOOR | 所在楼层的编号 | an..10 | 选填 | 汉字、字母、数字或其组合，例如1、2、B1等 |
|  | 楼层代码 | FLOORCODE | 楼层的顺序编码 | n3 | 选填 | GB/T 39609中的“楼层代码” |
|  | 户室 | ROOM | 所在户室的编号 | an..10 | 选填 | 汉字、字母、数字或其组合，例如101、204等 |
|  | 户室代码 | ROOMCODE | 户室的顺序编码 | n4 | 选填 | GB/T 39609中的“户室代码” |
|  | 网格 | GRID | 所在网格的名称 | an..50 | 选填 | 按照政法委网格区域划分 |
|  | 网格编码 | GRIDCODE | 所在网格的代码 | n3 | 选填 | 按照政法委网格编码规则编制 |
|  | 网格属性 | GPROPERTY | 网格属性的类型名称 | a..10 | 选填 |  |
|  | 网格属性代码 | PROERTYCODE | 网格属性的代码 | n2 | 选填 | 01—城市02—乡村 |
|  | 派出所 | POLICE | 所在行政区划下的派出所名称 | an..50 | 选填 |  |
|  | 派出所代码 | POLICECODE | 派出所的代码 | n3 | 选填 | 按照公安片区编码规则编制 |
|  | 警务区 | PDISTRICT | 派出所下的警务区的名称 | an..50 | 选填 |  |
|  | 警务区代码 | PDCODE | 警务区的代码 | n3 | 选填 | 按照公安片区编码规则编制 |
|  | 专业区 | SZONE | 所在矿区、林区、牧区、自然保护区、开发区、工业区、保税区的名称 | an..48 | 选填 |  |
|  | 组 | SQUAD | 所在村民小组的名称 | an..36 | 选填 |  |
|  | 地址简称 | SHORTNAME | 地名地址的简单称谓 | an..100 | 必填 | 汉字、字母、数字和符号组合，简称如有多个，以英文半角分号（“；”）隔开 |
|  | 地址全称 | NAME | 地名地址的详细名称 | an..100 | 必填 |  |
|  | 地址俗称 | ALIASNAME | 地名的其他通俗名称 | an..100 | 必填 | 汉字、字母、数字和符号组合,俗称如有多个，以英文半角分号（“;”）隔开 |
|  | 地址编码 | GUID | 每条地名地址数据记录的不重复的编码 | n41 | 必填 | 不可重复，不可修改 |
|  | 坐落描述 | LOCADES | 建筑物的所在位置的描述信息 | an..100 | 选填 |  |
|  | 检索词 | SEARCHTERM | 检索地址的关键词 | an..50 | 选填 | 检索词如有多个，以英文半角分号（“;”）隔开 |
|  | 图片 | PICTURE | 图片文件路径 | an..500 | 选填 |  |
|  | 经度 | LONGITUDE | 地名地址的经度位置 | n..13,10 | 选填 | 单位为度 |
|  | 纬度 | LATITUDE | 地名地址的纬度位置 | n..13,10 | 选填 | 单位为度 |
|  | 高程 | ALTITUDE | 地名地址的高程信息 | n..13,10 | 选填 | 单位为米 |
|  | 房屋类型 | BUILDINGTP | 房屋类型的标签 | an..20 | 选填 | 房屋类型如有多个，以英文半角分号（“;”）隔开 |
|  | 备注 | REMARK | 有关地名地址的其他说明信息 | ..ul | 选填 |  |

* 1. 数据处理
		1. 概述

通过共享平台采集不同来源部门已有地名地址数据时，应对地址数据进行数据清洗、比对、去重等处理操作后批量入库。通过采集系统采集的地名地址数据，应经过数据处理和检测后统一入库，针对问题数据应及时反馈给数据提供部门进行整改。

* + 1. 数据清洗

地名地址数据格式不符合要求，或含有格式要求外的字符时，应通过违规字符清除、违规字符替换等操作进行处理，包括但不限于以下要求：

1. 违规字符清除：对于存在违规字符的数据，应通过设置规则清理数据字段中的违规字符，保留合规字符。包括但不限于以下要求：
	1. 地名地址数据中含有如“￥”等违规字符；
	2. 省、市、区县等字段中含有空格；
	3. 地址编码等字段中含有英文字母或字符。
2. 违规字符替换：对含有违规字符的内容进行字符替换，将违规字符替换为合规字符，包括但不限于以下要求：
	1. 统一替换：将违规字符统一替换为某合规字符；
	2. 规则替换：针对不同违规字符对应替换不同的合规字符；
	3. 精确替换：设置匹配规则做精确替换，比如“&”替换为“和”；
	4. 模糊替换：设置匹配规则做模糊替换，比如“&\*\*\*（\*代表其他字符）”替换为“和”。
3. 重复字符去重：对于存在重复字符的地址数据，应通过设置规则去除重复内容，保留关键字符，比如，“文登区香山南路3-1-3号文登区”修改为“文登区香山南路3-1-3号”。
	* 1. 数据比对

针对多个来源的数据项，应根据数据业务场景分析，约定比对和校验规则，并根据预定义的规则，对数据进行比对处理，并将结果反馈数据提供部门，确保入库前的数据质量。包括但不限于以下要求：

1. “一数一源”策略：同一数据来自多个部门时，以权威部门来源为准，比如“警务片区”以公安部门提供为准，“网格数据”以政法委为准；
2. 大多数原则策略：同一数据存在多条不同记录时，以大多数记录为准，比如某小区的“地址”数据，三条记录为“威海市文登区香山南路”，一条记录为“威海市文登区香山北路”，取“威海市文登区香山南路”为其地址信息；
3. 新鲜度原则策略：同一数据存在不同的时间属性时，以时间最近的数据为准，比如某人的“家庭住址”信息，两年前为“幸福小区”，两年后为“花园小区”，取“花园小区”为其家庭住址信息。
	* 1. 数据标准化

对于因历史问题或格式问题存储的数据，可通过相关数值标准化规则进行处理，包括但不限于以下要求：

1. 地名地址数据中的小写字母应转换为大写字母，比如，“环翠区凤林小区a区”转换为“环翠区凤林小区A区”；
2. 地名地址数据中的“\_”应转换为“-”，比如，“文登区香山南路3\_1\_3号”转换为“文登区香山南路3-1-3号”；
3. 地名地址数据中的全角字符应转化为半角字符；
4. 地名地址数据的格式不符合要求时，应根据相关业务说明抽取地址相关字段并转化为内部统一地址规格，并建立原始地址和内部地址的关联关系，比如，“山东省环翠区威海市第一中学”转换为“山东省威海市环翠区文化中路75号威海市第一中学”。
	1. 数据入库
		1. 入库流程

数据入库包括数据库建立、入库前检查、数据导入、入库后检查四个流程。

* + 1. 入库前检查

数据入库前应开展检查，包括但不限于以下要求：

1. 入库数据应符合5.3的要求；
2. 应重点检查数据属性是否完备，数据属性取值是否合法。对象内部的数据属性是否一致，是否符合规范要求；
3. 应重点检查坐标是否准确，地址之间的空间关系是否合理；
4. 应重点检查要素之间的关系是否合理；
5. 结构化描述中各组成部分应与属性字段市、县（市、区）、街道（乡镇）、村居（社区）、专业区、街路巷、组、自然村、小区、门址、楼址、单元、楼层、户室内容保持一致；
6. 市、县（市、区）街道（乡镇）、村居（社区）、专业区、街路巷、组、自然村、小区等属性字段，地名地址完全落入时应全部采集入库。
	* 1. 数据导入

地名地址数据宜采用手工输入、批量或自动入库等方式入库，其中自动入库采用服务接口的方式实现。入库后应记录数据入库日志。

* + 1. 入库后检查

数据入库后应进行检查，对发现问题及时形成入库日志，入库后检查内容包括但不限于：

1. 数据完整性：是否存在重复数据；
2. 数据一致性：入库前后图形、属性、标识码是否保持一致；
3. 数据唯一性：地名地址编码信息是否保持唯一。
	1. 数据更新
		1. 更新方式

数据更新的方式主要包括：

1. 主动更新：网格员根据日常维护需求，主动更新所在辖区的地名地址数据；
2. 被动更新：在地名地址应用过程中行政区划变更或收到数据使用方反馈问题时，对地名地址数据核查更新。
	* 1. 更新要求

应首先对原始库数据进行更新，然后通过平台抽取至标准库，实现同步更新，数据更新包括但不限于以下要求：

1. 数据应以增量方式进行更新，并对更新的数据进行记录，用于数据溯源；
2. 地名地址编码应根据地名地址变化而同步更新。
	1. 数据安全

统一地址库数据安全要求应符合GB/T 22239—2019、GB/T 35273—2020、DB37/T 3523.2—2019中的规定。

1.

参考文献

[1] GB/T 39609—2020 地名地址地理编码规则

[2] GA/T 2000.3—2014 公安信息代码　第3部分：治安管理信息数据归属单位编码规则

[3] CJ/T 349—2010 数字社区管理与服务 网格划分与编码规则

[4] 国土资源部,国家测绘地理信息局.关于加快使用2000国家大地坐标系的通知: 国土资发〔2017〕

30号[A/OL].(2017-03-10)[2017-03-27].http://www.mnr.gov.cn/gk/tzgg/201712/t20171

212\_1992109.html.

[5] 国务院.关于启用“1985国家高程基准”的批复:国函〔1987〕78号[A/OL].(1987-05-16)[1987

-05-16].http://www.gov.cn/xxgk/pub/govpublic/mrlm/201103/t20110330\_63783.html.

