

DB 3203

徐 州 市 地 方 标 准

DB 3203/T 1056—2024

数据元件 第 1 部分：结构要求

Data components—Part 1: Requirements for structure

2024 - 07 - 31 发布

2024 - 09 - 01 实施

徐州市市场监督管理局 发 布

目 次

前言 III

引言 IV

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 结构 2

5 类型 2

 5.1 总体原则 2

 5.2 组态数据元件 3

 5.3 模态数据元件 3

 5.4 组合态数据元件 3

6 命名 3

 6.1 要素 3

 6.2 方式 3

7 标识 4

 7.1 组成形式 4

 7.2 世界各国和地区名称代码 4

 7.3 区域码 4

 7.4 类别码 4

 7.5 类型码 5

 7.6 顺序码 5

 7.7 子码 5

 7.8 应用示例 5

8 核心元数据 5

 8.1 属性 5

 8.1.1 取值 5

 8.1.2 中文名称 5

 8.1.3 定义 5

 8.1.4 英文名称 5

 8.1.5 短名 6

 8.1.6 数据类型 6

 8.1.7 约束类型 6

 8.1.8 取值示例 6

 8.2 描述 6

 8.2.1 业务元数据 6

 8.2.2 技术元数据 7

8.2.3 管理元数据..... 9

附录 A （资料性）数据资源到数据元件的转换过程及应用场景示例 17

参考文献..... 19

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。

本文件是DB 3203/T 1056《数据元件》的第1部分。DB 3203/T 1056已经发布了以下部分：

——第1部分：结构要求。

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。

本文件由徐州市大数据管理中心提出并归口。

本文件起草单位：徐州市大数据管理中心、中电数据产业集团有限公司、中电数创（北京）科技有限公司。

本文件主要起草人：谭睿、陆志鹏、王冬梅、国丽、胡成盛、许朝、孙晓冰、李宏辉、杨二龙、韩耀、刘松江、黄岩、孙自立、杨伟伟。

引 言

数据元件是连接数据供需两端的“中间态”，是原始数据与应用之间的数据初级产品和交易标的物，实现了从设计、开发调试、生产管理到入库编目的全流程安全合规管控，进而实现“数据可用不可见，数据不动程序动”。

DB 3203/T 1056《数据元件》是结合地方法规要求，指导数据元件开发商和数据运营商进行数据元件的设计、开发、交易和管理。拟由五个部分组成：

- 第1部分：结构要求。目的在于对数据元件的结构进行规范化表示，对数据元件设计、开发等提供指导，并给出相关信息；
- 第2部分：安全审核要求。目的在于对数据元件的设计、开发等流程进行安全审核，保障对数据元件的安全管控，对数据元件的审核流程提供指导，并给出相关信息；
- 第3部分：质量评价与管理规范。目的在于对数据元件的质量进行规范化评价和管理，对数据元件的质量检测流程提供指导，并给出相关信息；
- 第4部分：分类分级指南。目的在于提升数据元件的安全管理水平，对数据运营商提供方法指导，并给出相关信息；
- 第5部分：估值与定价指南。目的在于对数据元件的价值进行合理评估和定价，指导数据元件开发商和数据运营商对数据元件进行价值评估，并给出相关信息。

本文件规定了数据元件的结构、类型、命名、标识和核心元数据的要求，明确了数据元件的规范化表示，对数据元件开发商和数据运营商开发和管理数据元件具有指导意义。

数据元件 第1部分：结构要求

1 范围

本文件规定了数据元件的结构、类型、命名、标识和核心元数据的要求。

本文件适用于数据要素生产和流通过程中数据元件的描述、设计、开发与应用，所定义的数据元件类型主要为结构化数据，不涉及自然语言、图像等非结构化与其他复杂类数据。

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 2260 中华人民共和国行政区划代码

GB/T 2659.2 世界各国和地区及其行政区划名称代码 第2部分：行政区划代码

GB/T 21063.3 政务信息资源目录体系 第3部分：核心元数据

DB 32/T 4040.1 政务大数据 数据元规范 第1部分：总则

3 术语和定义

下列术语和定义适用于本文件。

3.1

数据 data

任何以电子或者其他方式对信息的记录。

[来源：GB/T 41479—2022，3.1]

3.2

数据集 dataset

具有一定主题，可以标识并可以被计算机处理的数据集合。

[来源：GB/T 36344—2018，2.6]

3.3

数据资源 data resource

在生产、生活和管理等社会活动中以数字化形式留存的数据总称。

3.4

数据脱敏 data desensitization

对数据资源进行处理，屏蔽敏感数据的一种数据保护方法。

3.5

数据元件 data component

具有一定主题，通过对数据资源脱敏处理后，根据需要由若干相关字段形成的数据集或由数据资源的关联字段通过建模形成的数据特征。

3.6

组态数据元件 composed data component

数据资源脱敏处理后，由若干个相关字段形成的数据集。

3.7

模态数据元件 modal data component

数据资源的关联字段通过建模形成的数据特征。

3.8

组合态数据元件 combinatorial data component

组态数据元件和模态数据元件的组合。

3.9

数据元件模型 data component model

使用数据资源开发数据元件过程中所采用的某些规则、算法等构造的数据模型。

3.10

元数据 metadata

定义和描述其他数据的数据。

[来源：GB/T 26816—2011，3.2.18]

4 结构

4.1 数据元件的结构内容应包括以下四部分：

- a) 类型：从形态特征来进行描述；
- b) 命名：从命名要素和命名方式来描述；
- c) 标识：从标识符组成和标识符形式来描述；
- d) 核心元数据：包括业务元数据、技术元数据和管理元数据。

4.2 数据元件的结构见图 1。

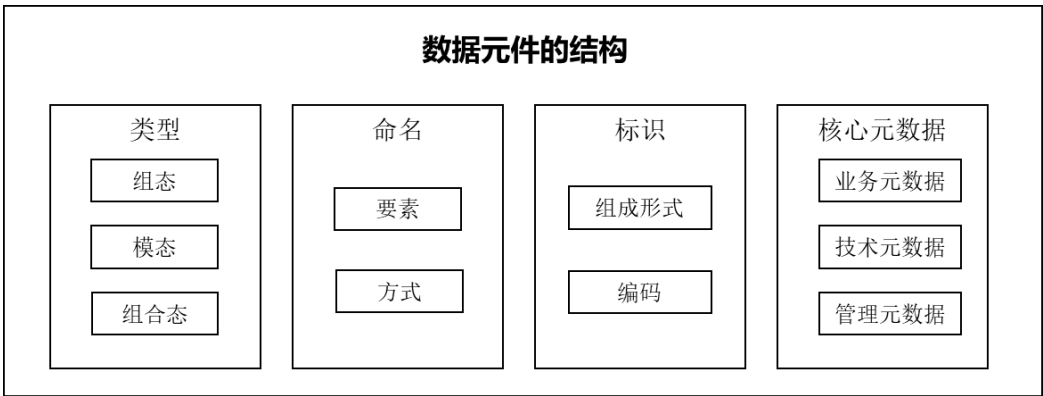


图1 数据元件的结构图

5 类型

5.1 总体原则

5.1.1 数据元件应遵守以下原则：

- a) 最小必要原则：数据元件构建所使用的数据资源不宜包含过多信息，对数据资源表的字段数、记录数和数据存储量等进行约束控制；

- b) 去标识化或主体授权同意原则：对于涉及政府、组织、企业或个人等重要数据和敏感信息，去除主体标识或者主体授权同意；
- c) 不可逆原则：数据元件模型保证数据元件结果不可逆回到数据资源，确保数据的使用者无法根据数据元件提供的信息还原出数据原貌。

5.1.2 数据资源到数据元件的转换过程及应用场景示例，见附录 A。

5.2 组态数据元件

5.2.1 组态数据元件应对原始数据资源中敏感信息进行脱敏处理，除敏感字段外，输出字段内容与选取的原始数据资源字段相同。

5.2.2 原始数据资源使用的有效数据字段个数应不少于 3 个、应不超过 12 个。

5.2.3 单次输出的数据量应不超过 30 万条记录。

示例：企业基本信息包含企业名称、企业人员规模、企业营业收入等，通过对企业名称进行脱敏处理后，形成徐州市企业基本信息元件，表示为：2022 年徐州市规上企业基本信息元件。

5.3 模态数据元件

5.3.1 模态数据元件输出的特征字段应是新的字段。

5.3.2 模态数据元件包含的特征字段个数应为 1 个。

5.3.3 单个模态数据元件应是单个对象输出的特征记录。

5.3.4 同类模态数据元件单次输出的数据量应不超过 30 万条记录。

示例：企业财务信息包括资产负债率、净资产、主营业务占比等字段信息，通过数学建模可形成财务特征元件，用于企业信贷业务，表示为：2022 年徐州市中小企业财务特征元件。

5.4 组合态数据元件

5.4.1 原始数据资源使用的有效数据字段个数应不少于 3 个、应不超过 12 个。

5.4.2 组合态数据元件包含的特征字段个数应不超过 12 个。

5.4.3 单次输出的数据量应不超过 30 万条记录。

示例：2022 年徐州市中小企业经营特征元件包含 2022 年徐州市中小企业基本信息和 2022 年徐州市中小企业财务特征，用于企业信贷业务，表示为：2022 年徐州市中小企业经营特征元件。

6 命名

6.1 要素

6.1.1 数据元件命名应包含以下要素：

- a) 限定要素：
 - 1) 时间：可选要素，表示数据元件涉及的时间范围，如年、月等；
 - 2) 地域：可选要素，表示数据元件涉及的地域，包括国家、地区等；
 - 3) 行业领域：可选要素，表示数据所属的行业领域。
- b) 主体要素：
 - 1) 数据对象：必选要素，表示数据元件所涉及的数据对象主体；
 - 2) 信息主题：必选要素，表示数据元件所包含内容的概要描述及数据粒度相关信息。

6.1.2 数据元件命名长度不宜超过 30 个字符。

6.2 方式

- 6.2.1 数据元件命名应符合“限定要素”+“主体要素”+“数据元件”表达方式的要求。
- 6.2.2 限定要素的取舍应符合以下要求：
- a) 数据元件的命名中不包含限定要素时，无其描述；
 - b) 不造成歧义时，可省略其描述；
 - c) 命名要素不满足实际描述需求时，应根据数据元件内容增加时间、地域或行业领域的限定要素描述。

示例1：徐州市人流量分布数据元件（其中不含时间和行业领域限定要素）。

示例2：2022 年徐州火车站客流量统计数据元件（省略行业领域限定要素）。

7 标识

7.1 组成形式

标识应由数据元件的世界各国和地区名称代码、区域码、类别码、类型码、顺序码和子码依次连接组成，见图2。

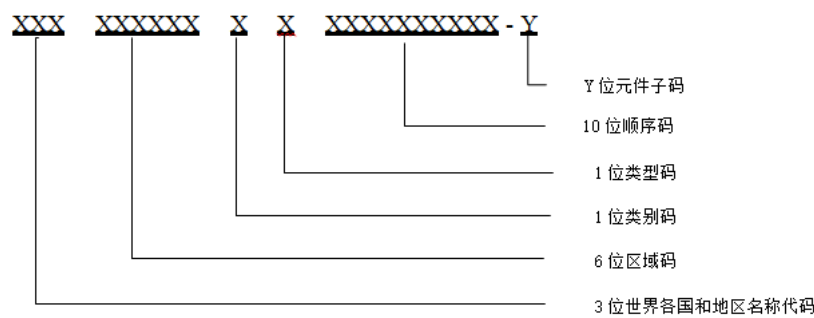


图2 数据元件的标识组成形式

7.2 世界各国和地区名称代码

7.2.1 世界各国和地区名称代码为数据资源所在国家或地区代码，应采用三位大写字母编码，应符合 GB/T 2659.2 中关于“三字符拉丁字母代码结构”的规定。

示例：中国 CHN

7.2.2 数据源不属于单个国家或地区时，则应用大写字母 AAA 表示，数据来源不明确，则应用大写字母 XXX 表示。

7.3 区域码

7.3.1 数据资源归属地（省、市、县）的行政区划代码，应采用 6 位数字代码，应符合 GB/T 2260 的规定。

示例：徐州 320300

7.3.2 数据资源包括两个或两个以上行政区，应采用上一级别行政区代码表示；当上一级别为国家时，应用 111111 表示。

7.3.3 数据来源不明确时，则应用 000000 表示，数据资源为境外资源时，则应用 999999 表示。

7.4 类别码

类别码应由 1 位阿拉伯数字表示：

——1：标准数据元件；

——2：定制数据元件。

7.5 类型码

类型码应由1位阿拉伯数字表示：

——1：组态数据元件；

——2：模态数据元件；

——3：组合态数据元件。

7.6 顺序码

顺序码为数据元件的顺序编号，应由10位阿拉伯数字表示，其中前4位为年份，后6位为数据元件的设计或生产或制造的顺序编号。

7.7 子码

子码为同类数据元件自定义对象个体代号，应由不定长数字与字母组合。根据应用的需求应自定义数据元件子码的数量和编码的位数。

注：数据元件子码的数量和编码的位数由数据元件中对象个数测算得出。

7.8 应用示例

由江苏省徐州市云龙区的数据资源作为原材料加工，2022年生产的第123个准模态数据元件的1号数据元件主体对象的记录表示为：CHN320303122022000123-1。

由境外（美国）的数据资源作为原材料加工，2022年生产的第123个定制组态数据元件表示为：USA999999212022000123。

8 核心元数据

8.1 属性

8.1.1 取值

数据元件的核心元数据元素应按照以下方式进行取值：

——中文名称；

——定义；

——英文名称；

——短名；

——数据类型；

——约束类型。

8.1.2 中文名称

以字符串的形式表示。

8.1.3 定义

以字符串的形式表示。

8.1.4 英文名称

以字符串形式表示。

8.1.5 短名

应符合 GB/T 21063.3 的取值规则。

8.1.6 数据类型

数据类型应符合下列五种：

- a) 字符型：描述字符类型的属性；
- b) 数值型：描述整数、浮点数等类型的属性；
- c) 日期型：描述有日期相关的属性；
- d) 日期时间型：描述与日期和时间相关的属性；
- e) 布尔型：描述是/否、真/假等类型的属性。

8.1.7 约束类型

约束类型应符合下列三种：

- a) M：必选，表示该数据元件属性是必备的；
- b) C：条件必选，满足约束条件中所规定的要求时必须选择；
- c) O：可选，表示该数据元件属性根据实际应用是可选的。

8.1.8 取值示例

数据元件的核心元数据元素开发单位名称的取值为：XX信息技术有限公司。

8.2 描述

8.2.1 业务元数据

8.2.1.1 取值

数据元件的业务元数据取值应包括：

- 标识符；
- 名称；
- 英文名；
- 类型；
- 摘要；
- 关键词。

8.2.1.2 标识符

定 义：在特定范围内给予数据元件的唯一标识。按照第7章给出的规则编制。

英文名称：component ID。

短 名：comID。

数据类型：字符型。

约束类型：M。

取值示例：CHN320300122022000321-1。

8.2.1.3 名称

定 义：数据元件的命名。按照第6章给出的规则编制。

英文名称：componentTitle。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/906143212114010215>