ICS 07.080
A40

T/SZGIA

团 体 标 准

T/SZGIA 6.1—2019

基因检测产品数据标准
第1部分：通用标准

Genomics Data Normalization –

Part 1: General Specification

2019－06－21 发布　　2019－06－30 实施

深圳基因产学研资联盟　发布

目　　次

前　　言

《基因检测产品数据标准》包括通用标准和特定检测产品的数据标准，如：

——第1部分：通用标准；

——第2部分：孕妇外周血胎儿游离DNA产前检测元数据目录；

本标准按照GB/T 1.1-2009给出的规则起草。

本部分起草单位：深圳华大基因科技有限公司、深圳华大生命科学研究院、深圳华大临床检验中心、深圳华大基因股份公司、深圳基因产学研资联盟、北京诺禾致源科技股份有限公司、广州医科大学附属第三医院、深圳瑞奥康晨生物科技有限公司、菁良基因科技（深圳）有限公司、深圳裕策生物科技有限公司。

本部分主要起草人：吕春杰、刘小燕、唐美芳、李陶莎、程奇、李倩一、吴昊、李瑞强、吴俊、王大伟、黎青、陈敏、郑晨晴、杨旭、饶颖、李淼、聂新华、高志博。

引　　言

组学数据的数据类型可分为非结构化数据和结构化数据。其中非结构化数据，通过数据格式、数据格式规格说明、归档目录描述；结构化数据，通过数据元、值域来描述。

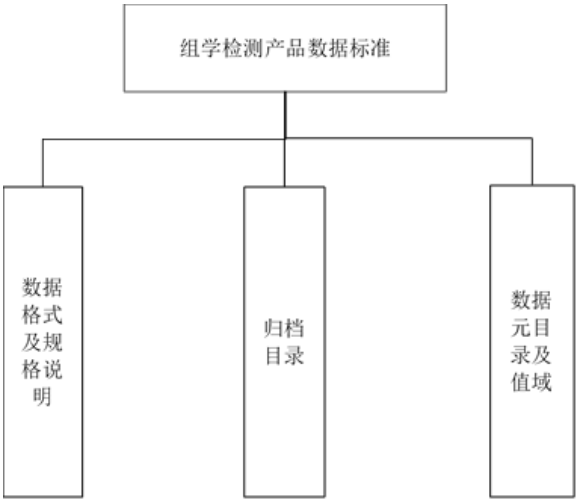

图1　组学产品数据标准框架

数据格式及规则说明规范了数据格式的内容结构、属性与描述规则、数据格式的编制规则。

归档目录规范了归档目录的内容结构、属性与描述规则、格式和索引的编制规则。

数据元规范化定义了数据元的编码方法及描述属性，描述属性包括状态、来源、基础数据集、基础数据元标识符、数据元中文名称、定义、必要性、信息保护、数据元的数据类型、表示格式、单位代码、数据元允许值等。值域代码规范了数据元值域的编码方法、代码表格式和表示要求、代码表的命名与标识。

基因检测产品数据标准
第 1 部分：通用标准

1　范围

本部分规定了基因组学产品数据属性与描述规则、数据元索引与数据元值域的编码方法、代码表格式与表示要求、代码表的命名与标识、数据格式的内容结构与编制规则、数据格式的规格说明、归档目录结构的内容。

本文件适用于基因组学产品数据标准的编制。

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 2312 信启、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 10113 分类与编码通用术语

GB/T 17295 国际贸易用计量单位代码

GB/T 18391.1 信息技术 元数据注册系统（MDR） 第1部分：框架

GB/T 18391.信息技术 元数据注册系统（MDR） 第3部分：注册系统元模型和基本属性

GB/T 19488.1 电子政务数据元 第1部分：设计和管理规范

WS/T 303 卫生信息数据元标准化规则

WS/T 305 卫生信息数据集元数据规范

WS/T 306 卫生信息数据集分类与编码规则

WS 363.1 卫生信息数据元目录 第1部分：总则

WS 364.1 卫生信息数据元值域代码 第1部分：总则

JT/T 697.1 交通信息基础数据元 第1部分：总则

3　术语和定义

3.1

数据元标识符 data element identifier

数据元目录中为数据元分配的与语言无关的唯一标识。

注：该定义源于国家卫生行业标准WS363.1中3.1。

3.2

数据元公用属性 public attribute

在数据元目录中数据元的属性值均相同的属性。如本标准中注册机构。

注：该定义源于国家卫生行业标准WS363.1中3.2。

3.3

数据元专用属性 specialized attribute

在数据元目录中数据元属性值不相同的属性。

注：该定义源于国家卫生行业标准WS363.1中3.3。

3.4

值域 value domain

允许值的集合。

注：该定义源于国家标准GB/T 18391.1中的3.75。

3.5

类别 category

具有某种共同属性（或特征）的事物（或概念）的集合。

注：该定义源于国家标准GB/T 10113中的2.1.1。

3.6

分类 classification

按照选定的属性（或特征）区分分类对象，并将具有某种共同属性（或特征）的分类对象集合在一起的过程。

注：该定义源于国家卫生行业标准WS364.1中3.3。

3.7

线分类法 method of line classification

将分类对象按选定的若干属性（或特征）逐次地分为若干层级，每个层级又分为若干类目，不同层级类目之间构成隶属关系。这种分类方法称为线分类法。

注：该定义源于国家卫生行业标准WS364.1中3.4。

3.8

面分类法 method of area classification

选定分类对象的若干属性（或特征），将分类对象按每属性（或特征）划分成一组独立的类目，每一组类目构成一个“面”。再按一定顺序将各个“面”平行排列。使用时根据需要将有关“面”中的相应类目按“面”的指定排列顺序组配在一起，形成一个新的复合类目。这种分类方法称为面分类法。

注：该定义源于国家卫生行业标准WS364.1中3.5。

3.9

代码 code

表示特定事物（或概念）的一个或一组字符。这些字符可以是阿拉伯数字、拉丁字母或便于电子计算机和人识别与处理的其他符号。

注：该定义源于国家标准GB/T 10113中的2.2.5。

3.10

编码 coding

给事物（或概念）赋予代码的过程。

注：该定义源于国家卫生行业标准WS364.1中3.7。

3.11

代码结构 code structure

一个完整代码的组成方式和长度的综合表示。

注：该定义源于国家卫生行业标准WS364.1中3.8。

3.12

代码类型 code type

从某一个方面（如含义、结构、长度、组成等）来表示代码的某种特性。

如：从含义上可分为有含义代码和无含义代码；从结构上可分为层次码和顺序码等；从长度上可分为等长代码和不等长代码；从组成上可分为数字代码和字母代码等。

注：该定义源于国家卫生行业标准WS364.1中3.9。

3.13

无含义代码 unmeaning code

对编码对象只起标识作用，而无任何其他附加含义的代码。

注：该定义源于国家卫生行业标准WS364.1中3.10。

3.14

有含义代码 meaning code

除对编码对象起标识作用外，还具有其他特定含义的代码。

注：该定义源于国家卫生行业标准WS364.1中3.11。

3.15

数字型代码 numeric code

由阿拉伯数字（0~9）构成的代码。

注：此种类型的代码仅仅是以阿拉伯数字的形式表示，但不是数值型，不可直接用于计算。

注：该定义源于国家卫生行业标准WS364.1中3.12。

3.16

字母型代码 alphabetic code

由字母构成的代码。

注：其中所称字母通常为英文字母（I、O因与1、0相似，通常不使用）。

注：该定义源于国家卫生行业标准WS364.1中3.13。

3.17

字母数字型代码 alphanumeric code

由字母和阿拉伯数字混合构成的代码。

注：该定义源于国家卫生行业标准WS364.1中3.14。

3.18

层次码 layer code

以编码对象的隶属关系为排列顺序而组成的有层级关系的代码。

注：该定义源于国家卫生行业标准WS364.1中3.15。

3.19

顺序码 sequential code

按照阿拉伯数字或字母的自然顺序来表示编码对象的代码。亦称“流水码”。

注1：通常情况下，顺序码是连续的，代码之间不出现断点。但在特殊情况下，可采用等距离（间隔）跳跃式编码。

注2：该定义源于国家卫生行业标准WS364.1中3.16。

3.20

系列顺序码 alignment-sequence code

根据编码对象属性（或特征）的相同或相似，将编码对象分为若干组。再将顺序码分为相应的若干系列（也称为“段”），并分别赋给各编码对象组。在同一系列内对编码对象连续编码，并预留扩展空间。这样编制的代码称为系列顺序码。

注：该定义源于国家卫生行业标准WS364.1中3.17。

3.21

等长代码 code of equal length

在同一个代码体系中，所有编码对象的代码长度都相等。

注：该定义源于国家卫生行业标准WS364.1中3.18。

3.22

不等长代码 code of different length

在一个完整的代码体系中，代码总长度不完全相同。

注：该定义源于国家卫生行业标准WS364.1中3.19。

3.23

标识符 identifier

在特定语境中，可唯一性地标识与之相关联的事物的一系列字符，可看做用来识别特定对象的数据元的编码值。

注：该定义源于国家卫生行业标准WS364.1中3.20。

3.24

归档目录标识符 directory identifier

归档目录结构中为归档目录分配的与语言无关的唯一标志。

3.25

归档目录公用属性 public attribute

在归档目录结构中数据目录的属性值均相同的属性。如本标准中注册机构。

3.26

归档目录专用属性 specialized attribute

在归档目录结构中数据目录的属性值不相同的属性。

4 缩略语

DE 数据元（Data Element）；
DI 数据标识符（Data Identifier）；
DNA 脱氧核糖核酸（DeoxyriboNucleic Acid）；
RO 主管机构（Responsible Organization）；
RA 注册机构（Registration Authority）；
RAI 注册机构标识符（Registration Authority Identifier）；
SO 提交机构（Submitting Organization）；
VI 版本标识符（Version Identifier）。

5 数据格式属性与描述规则

数据格式属性设置参照WS/T 303，统一规定采用5类14项属性，并按通用性程度分为两类：数据元公用属性和数据元专用属性。数据元公用属性包括7项，数据元专用属性包括7项，见表1。

表1 数据元属性

序号	属性种类	数据元属性名称	约束	备注
1	标识类	数据格式标识符	必选	专用属性
2		数据格式名称	必选	专用属性
4		版本	必选	共用属性
5		注册机构	必选	共用属性
6		相关环境	必选	共用属性
7	定义类	适用范围	必选	专用属性
8	关系类	分类模式	必选	共用属性
11	表示类	数据格式允许值	必选	专用属性
12	管理类	主管机构	必选	共用属性
13		注册状态	必选	共用属性
14		提交机构	必选	共用属性

5.1 数据格式属性描述规则

5.1.1 数据格式标识符

数据元（DF）标识符采用字母数字混合码，包含数据标识符（DI）和版本标识符（VI）两级结构。

示例1：DI_V1

a） DI按照分类法和流水号相结合的方式，采用字母数字混合码。按照数据元对应的主题分类代码、大类代码、小类代码、顺序码、附加码从左向右顺序排列。其中：

——主题分类代码：用2位大写英文字母表示。代码统一定为“DF”。

——大类代码：用2位数字表示，数字大小无含义。

——小类代码：用2位数字表示，数字大小无含义；无小类时则小类代码为00。小类与大类代码之间加“.”区分。

——顺序码：用3位数字表示，代表某一小类下的数据元序号，数字大小无含义；从001开始顺序编码。顺序码与小类代码之间加“.”区分。

b）VI结构由4部分组成，为“V”+“m..m”+“.”+“n..n”。其中“m..m”和“n..n”为两位阿拉伯数字构成，在数学上应是具有意义的正整数。“m..m”表示主版本号，“n..n”表示次版本号。

示例2：“V1.2” 表示主版本为第一版，次版本为“第二版”。

如果数据元更新前后可以进行有效的数据交换，则更新后主版本号不变，次版本号等于当前次版本号加1；如果数据元更新前后无法进行有效的数据交换，则更新后主版本号等于当前主版本号加1，次版本号归0。

数据标识符结构见图2。

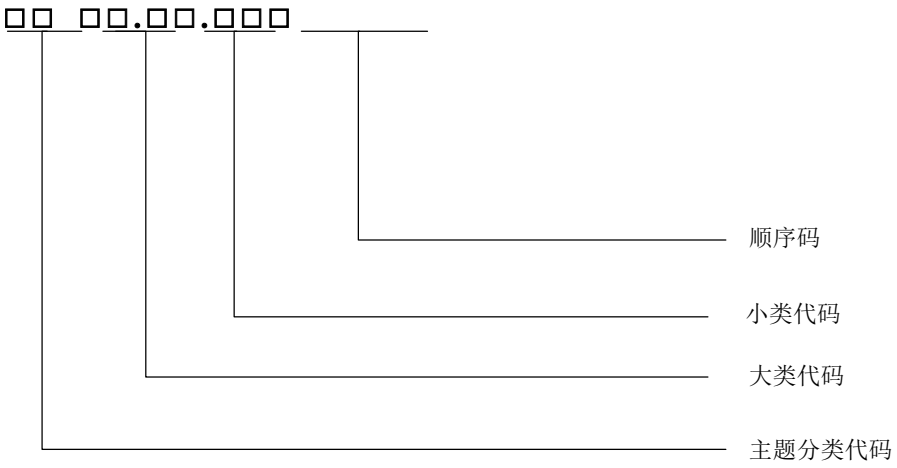

图2 数据标识符（DI）结构

5.1.2 数据格式名称

数据元“中文名称”应当是唯一的，并且以字母、汉字、数字式的字符串形式表示。

数据元的命名应使用一定的逻辑结构和通用的术语。

完整的数据元名称=对象类术语+特性类术语+表示类术语+（限定类术语）。

其中：

—— —个数据元需要有一个且仅有一个对象类术语。在组学数据元目录中若对象类术语为“本人”，则可酌情省略。

—— —个数据元需要有一个且仅有一个特性类术语。特性类术语是任何一个数据元名称所必需的成分，在数据元概念可以完整、准确、无歧义表达的情况下，其他术语可以酌情简略。

—— —个数据元需要有一个且仅有一个表示类术语。当表示类术语与特性类术语有重复或部分重复时，可从名称中将冗余词删除。通用表示类术语见表2。

—— 限定类术语由专业领域给定。限定类术语是可选的。

表2　通用表示类术语

表示词	含义
名称	表示一个对象称谓的一个词或短语
代码	替代某一特定信息的一个有内在规则的字符串（字母、数字、符号）
说明	表示描述对象信息的一段文字
金额	以货币为表示单位的数量，通常与货币类型有关
数量	非货币单位数量，通常与计量单位有关。计量单位参见附录表A.1，法定构成十进倍数和分数单位的词头见附录表A.2
日期	以公元纪年方式表达的年、月、日的组合
时间	以24小时制计时方式表达的—天中的小时、分、秒的组合
日期时间	完整时间表达格式，即DT15，YYYYMMDDThhmmss的格式
百分比	具有相同计量单位的两个值之间的百分数形式的比率
比率	一个计量的量或金额与另一个计量的量或金额的比
标志	又称指示符，两个且只有两个表明条件的值，如：是/否、有/无等
时长	两个时点间的时间长度

5.1.3　适用范围

本文件中数据格式适用范围以字母、汉字、数字式的字符串形式表示。

6　数据格式说明的编码方式

数据元（SF）标识符采用字母数字混合码，包含数据标识符（DI）和版本标识符（VI）两级结构。

示例1：DI_V1

c） DI按照分类法和流水号相结合的方式，采用字母数字混合码。按照数据元对应的主题分类代码、大类代码、小类代码、顺序码、附加码从左向右顺序排列。其中：

——主题分类代码：用2位大写英文字母表示。代码统一定为“SF”。

——大类代码：用2位数字表示，数字大小无含义。

——小类代码：用2位数字表示，数字大小无含义；无小类时则小类代码为00。小类与大类代码之间加“.”区分。

——顺序码：用3位数字表示，代表某一小类下的数据元序号，数字大小无含义；从001开始顺序编码。顺序码与小类代码之间加“.”区分。

d）VI结构由4部分组成，为“V”+“m..m”+“.”+“n..n”。其中“m..m”和“n..n”为两位阿拉伯数字构成，在数学上应是具有意义的正整数。“m..m”表示主版本号，“n..n”表示次版本号。

示例2：“V1.2” 表示主版本为第一版，次版本为“第二版”。

如果数据元更新前后可以进行有效的数据交换，则更新后主版本号不变，次版本号等于当前次版本号加1；如果数据元更新前后无法进行有效的数据交换，则更新后主版本号等于当前主版本号加1，次版本号归0。

数据标识符结构见图3。

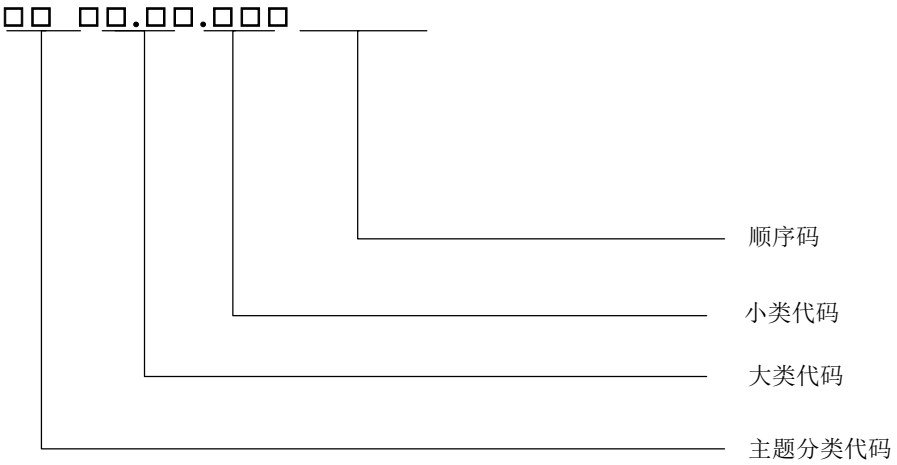

图3　数据标识符（DI）结构

7　归档目录属性及描述规则

数据格式属性设置参照WS/T 303，统一规定采用5类14项属性，并按通用性程度分为两类：数据元公用属性和数据元专用属性。数据元公用属性包括7项，数据元专用属性包括7项，见表3。

表3　归档目录属性

序号	属性种类	数据元属性名称	约束	备注
1	标识类	归档目录标识符	必选	专用属性
2		归档目录名称	必选	专用属性
4		版本	必选	共用属性
5		注册机构	必选	共用属性
6		相关环境	必选	共用属性
7	定义类	定义	必选	专用属性
8	关系类	分类模式	必选	共用属性
11	表示类	父目录	必选	专用属性
12	管理类	主管机构	必选	共用属性
13		注册状态	必选	共用属性
14		提交机构	必选	共用属性

7.1　归档目录属性描述规则

7.1.1　归档目录标识符

数据归档目录（FD）标识符采用字母数字混合码，包含数据标识符（DI）和版本标识符（VI）两级结构。

示例1：DI_V1

e） DI按照分类法和流水号相结合的方式，采用字母数字混合码。按照数据元对应的主题分类代码、大类代码、小类代码、顺序码、附加码从左向右顺序排列。其中：

——主题分类代码：用2位大写英文字母表示。代码统一定为“FD”。

——大类代码：用2位数字表示，数字大小无含义。

——小类代码：用2位数字表示，数字大小无含义；无小类时则小类代码为00。小类与大类代码之间加“.”区分。

——顺序码：用3位数字表示，代表某一小类下的数据元序号，数字大小无含义；从001开始顺序编码。顺序码与小类代码之间加“.”区分。

f）VI结构由4部分组成，为“V”+“m..m”+“.”+“n..n”。其中“m..m”和“n..n”为两位阿拉伯数字构成，在数学上应是具有意义的正整数。“m..m”表示主版本号，“n..n”表示次版本号。

示例2：“V1.2” 表示主版本为第一版，次版本为“第二版”。

如果数据元更新前后可以进行有效的数据交换，则更新后主版本号不变，次版本号等于当前次版本号加1；如果数据元更新前后无法进行有效的数据交换，则更新后主版本号等于当前主版本号加1，次版本号归0。

数据标识符结构见图4。

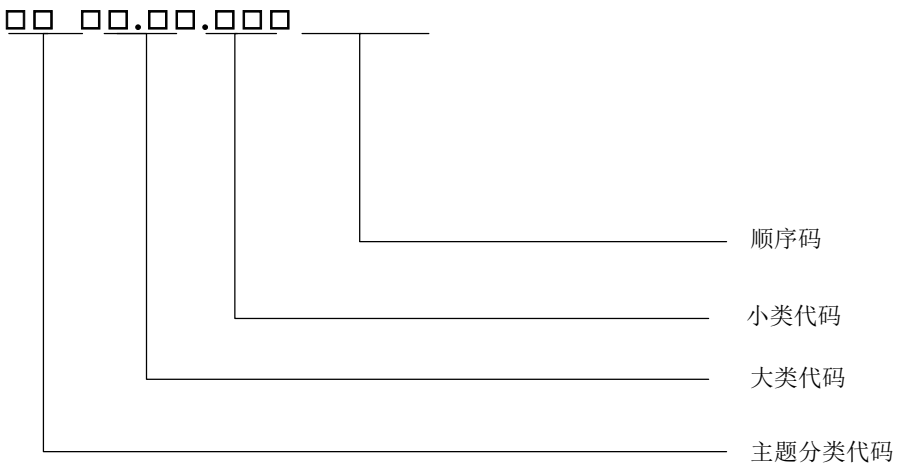

图4 数据标识符（DI）结构

7.1.2 归档目录名称

数据元“中文名称”应当是唯一的，并且以字母、汉字、数字式的字符串形式表示。

数据元的命名应使用一定的逻辑结构和通用的术语。

完整的数据元名称=对象类术语+特性类术语+表示类术语+（限定类术语）。

其中：

———个数据元需要有一个且仅有一个对象类术语。在组学数据元目录中若对象类术语为“本人”，则可酌情省略。

———个数据元需要有一个且仅有一个特性类术语。特性类术语是任何一个数据元名称所必需的成分，在数据元概念可以完整、准确、无歧义表达的情况下，其他术语可以酌情简略。

———个数据元需要有一个且仅有一个表示类术语。当表示类术语与特性类术语有重复或部分重复时，可从名称中将冗余词删除。通用表示类术语见表4。

——限定类术语由专业领域给定。限定类术语是可选的。

表4　通用表示类术语

表示词	含义
名称	表示一个对象称谓的一个词或短语
代码	替代某一特定信息的一个有内在规则的字符串（字母、数字、符号）
说明	表示描述对象信息的一段文字
金额	以货币为表示单位的数量，通常与货币类型有关
数量	非货币单位数量，通常与计量单位有关。计量单位参见附录表A.1，法定构成十进倍数和分数单位的词头见附录表A.2
日期	以公元纪年方式表达的年、月、日的组合
时间	以24小时制计时方式表达的—天中的小时、分、秒的组合
日期时间	完整时间表达格式，即DT15，YYYYMMDDThhmmss的格式
百分比	具有相同计量单位的两个值之间的百分数形式的比率
比率	一个计量的量或金额与另一个计量的量或金额的比
标志	又称指示符，两个且只有两个表明条件的值，如：是/否、有/无等
时长	两个时点间的时间长度

8　数据元属性与描述规则

8.1　数据元属性设置

数据元属性设置参照WS/T 303，统一规定采用5类14项属性，并按通用性程度分为两类：数据元公用属性和数据元专用属性。数据元公用属性包括7项，数据元专用属性包括7项，见表5。

表5　数据元属性

序号	属性种类	数据元属性名称	约束	备注
1	标识类	数据元标识符	必选	专用属性
2		数据元名称	必选	专用属性
3		信息保护	可选	专用属性
4		版本	必选	共用属性
5		注册机构	必选	共用属性
6		相关环境	必选	共用属性
7	定义类	定义	必选	专用属性
8	关系类	分类模式	必选	共用属性

9	表示类	数据元值的数据类型	必选	专用属性
10		表示格式	必选	专用属性
11		数据元允许值	必选	专用属性
12	管理类	主管机构	必选	共用属性
13		注册状态	必选	共用属性
14		提交机构	必选	共用属性

8.2 数据元属性描述规则

8.2.1 数据元标识符

数据元（DE）标识符采用字母数字混合码，包含数据标识符（DI）和版本标识符（VI）两级结构。

示例1：DI_V1

g） DI按照分类法和流水号相结合的方式，采用字母数字混合码。按照数据元对应的主题分类代码、大类代码、小类代码、顺序码、附加码从左向右顺序排列。其中：

——主题分类代码：用2位大写英文字母表示。代码统一定为“DE”。

——大类代码：用2位数字表示，数字大小无含义。

——小类代码：用2位数字表示，数字大小无含义；无小类时则小类代码为00。小类与大类代码之间加“.”区分。

——顺序码：用3位数字表示，代表某一小类下的数据元序号，数字大小无含义；从001开始顺序编码。顺序码与小类代码之间加“.”区分。

——附加码：用2位数字表示，代表一组数据元的连用关系编码；从01开始顺序编码，附加码与顺序号之间加“.”区分。无连用关系的数据元其附加码为“00”。

h）VI结构由4部分组成，为“V”+“m..m”+“.”+“n..n”。其中“m..m”和“n..n”为两位阿拉伯数字构成，在数学上应是具有意义的正整数。“m..m”表示主版本号，“n..n”表示次版本号。

示例2：“V1.2” 表示主版本为第一版，次版本为“第二版”。

如果数据元更新前后可以进行有效的数据交换，则更新后主版本号不变，次版本号等于当前次版本号加1；如果数据元更新前后无法进行有效的数据交换，则更新后主版本号等于当前主版本号加1，次版本号归0。

数据标识符结构见图5。

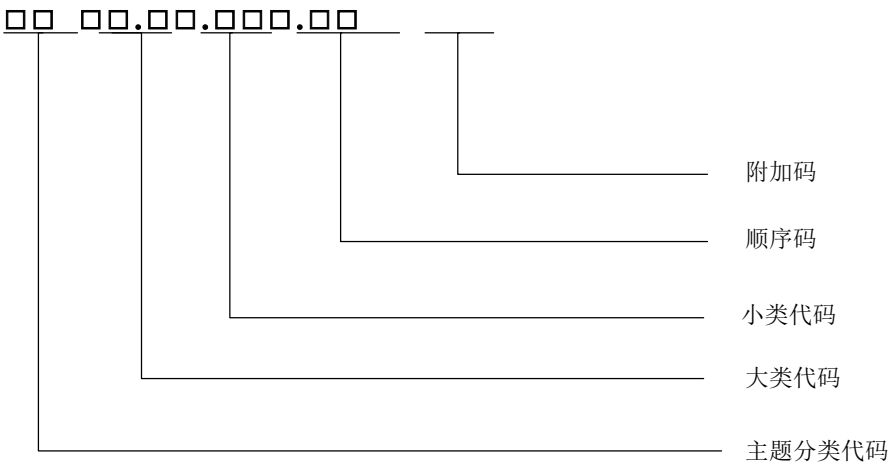

图5 数据标识符（DI）结构

8.2.2 数据元名称

数据元“中文名称”应当是唯一的，并且以字母、汉字、数字式的字符串形式表示。

数据元的命名应使用一定的逻辑结构和通用的术语。

完整的数据元名称=对象类术语+特性类术语+表示类术语+（限定类术语）。

其中：

———个数据元需要有一个且仅有一个对象类术语。在组学数据元目录中若对象类术语为“本人”，则可酌情省略。

———个数据元需要有一个且仅有一个特性类术语。特性类术语是任何一个数据元名称所必需的成分，在数据元概念可以完整、准确、无歧义表达的情况下，其他术语可以酌情简略。

———个数据元需要有一个且仅有一个表示类术语。当表示类术语与特性类术语有重复或部分重复时，可从名称中将冗余词删除。通用表示类术语见表6。

——限定类术语由专业领域给定。限定类术语是可选的。

表6　通用表示类术语

表示词	含义
名称	表示一个对象称谓的一个词或短语
代码	替代某一特定信息的一个有内在规则的字符串（字母、数字、符号）
说明	表示描述对象信息的一段文字
金额	以货币为表示单位的数量，通常与货币类型有关
数量	非货币单位数量，通常与计量单位有关。计量单位参见附录表A.1，法定构成十进倍数和分数单位的词头见附录表A.2
日期	以公元纪年方式表达的年、月、日的组合
时间	以24小时制计时方式表达的—天中的小时、分、秒的组合
日期时间	完整时间表达格式，即DT15，YYYYMMDDThhmmss的格式
百分比	具有相同计量单位的两个值之间的百分数形式的比率
比率	一个计量的量或金额与另一个计量的量或金额的比
标志	又称指示符，两个且只有两个表明条件的值，如：是/否、有/无等
时长	两个时点间的时间长度

8.2.3 定义

本文件中数据元定义以字母、汉字、数字式的字符串形式表示。

8.2.4 数据元值的数据类型

数据元值的数据类型描述规则见表3。本文件将字串型（S）分为三种形式，S1表示不可枚举的，且以字符描述的形式；S2表示枚举型，且列举值不超过3个；S3表示代码表的形式。

表7　数据元值的数据类型描述规则

数据类型	表示符	描述
字串型（string）	S	通过字符形式表达的值的类型。可包含字母字符（a~z，A~Z）、数字字符等。（默认GB 2312）
布尔型（boolean）	L	又称逻辑型，采用0（Fa1se）或1（True）形式表示的逻辑值的类型
数值型（number）	N	通过“0”到“9”数字形式表示的值的类型
日期型（date）	D	采用GB/T 7408中规定的YYYYMMDD格式表示的值的类型
日期时间型（datetime）	DT	采用GB/T 7408中规定的YYYYMMDDThhmmss格式表示的值的类型。（字符T作为时间的标志符，说明日的时间表示的开始。）
时间型（time）	T	采用GB/T 7408中规定的hhmmss格式表示的值的类型
二进制（binary）	BY	上述无法表示的其他数据类型，如图像、音频、视频等二进制流文件格式

8.2.5 表示格式

表示格式见表8和表9。

表8 数据元值的表示格式中字符含义描述规则

字符	含义
A	字母字符
N	数字字符
AN	字母或（和）数字字符
D8	采用YYYYMMDD的格式表示，其中，“YYYY”表示年份，“MM”表示月份，“DD” 表示日期
T6	采用hhmmss的格式表示，其中“hh”表示小时，“mm”表示分钟，“ss”表示秒
DT15	采用YYYYMMDDThhmmss的格式表示，字符T作为时间的标志符，说明日的时间表示的开始；其余字符表示与上同

表9 数据元值的表示格式中字符长度描述规则

类别	表示方法
固定长度	在数据类型表示符后直接给出字符长度的数目，如N4
可变长度	1）可变长度不超过定义的最大字符数 在数据类型表示符后加“..”后给出数据元最大字符数目，如AN..10 2）可变长度在定义的最小和最大字符数之间 在数据类型表示符后给出最小字符长度数后加“..”后再给出最大字符数，如AN4..20
有若干字符行表示的长度	按固定长度或可变长度的规定给出每行的字符长度数后加“X”后，再给出最大行数，如AN..40X3
有小数位	按固定长度或可变长度的规定给出字符长度数后，在“,”后给出小数位数，字符长度数包含整数位数、小数点位数和小数位数，如N6,2

应用示例：

示例1：S 字串型

AN10　　固定为10个字符（相当于5个汉字）长度的字符。

AN..10　可变长度，最大为10个字符长度的字符。

AN4..10　可变长度，最小为4个最大为10个字符长度的字符。

AN..20X3　可变长度，最多3行，每行最大长度为20个字符长度的字符。

示例2：N 数字型

N4　　固定长度为4位的数字。

N..4　　最大长度为4位的数字。

N6,2最大长度为6位的十进制小数格式（包括小数点），小数点后保留2位数字。

示例3：T 日期时间型

T8　　采用YYYYMMDD格式（8位定长）表示年月日。

T15 采用YYYYMMDDThhmmss格式（15位定长）表示年月日时分秒。时分秒之前加大写字母“T”。如2010年1月5日8时10分9秒为20100105T081009。

8.2.6 数据元允许值

本文件数据元值域有两种类型：

a)可枚举值域：由允许值列表规定的值域，每个允许值的值和值含义均应成对表示。其中：

——可选值较少的（如3个或以下），在“数据元允许值”属性中直接列举。

——可选值较多的（如3个以上），在“数据元允许值”属性中写出值域代码表名称。如代码表属引用标准的，则须注明标准号。

b)不可枚举值域：由描述规定的值域，在“数据元允许值”属性中须准确描述该值域的允许值。

9 数据元值域的编码方法

9.1 代码结构

数据元值域代码结构设计遵守以下要求：

a)代码结构设计注重代码的标识作用，避免承载过多的信息，以保证结构的简练。

b)代码结构符合信息处理的基本方法，尽量与系统内、外的相关标准结构协调一致。

c)代码结构确保代码的添加、删除和修改不破坏代码结构。

d)代码应采用便于使用的符号。

数据元值域代码结构描述遵守以下要求：

a)应明确描述所采用的代码种类、代码结构以及编码方法。

b)层次码的代码结构还可用示意图表示，如图6所示。

c)当代码结构复杂时，可用示例说明。

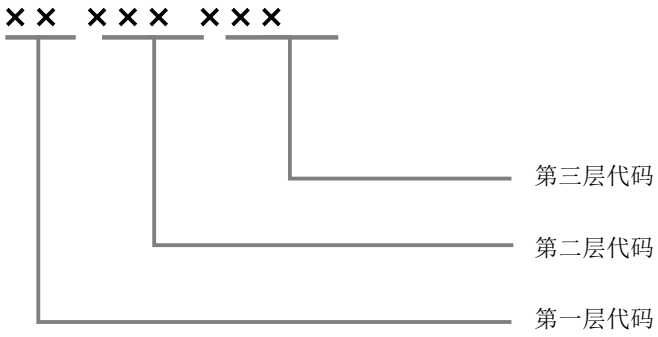

图6　数据元值域代码结构示意

9.2　代码长度

9.2.1　设计要求

数据元值域代码长度设计遵守以下要求：

a)在保证需求的前提下，代码长度应尽量简短。

b)尽可能使用等长代码，不宜使用不等长代码。

c)代码预留空间应满足编码对象的发展要求。

计算公式

代码长度的计算公式：

$$N=\sum_{i=1}^{n}\log_{a_i}Q_i \qquad \text{……（1）}$$

式中：

N——代码总长度；

n——层次码的层数；

i——第i层（或第i面）；

a_i——第i层（或第i面）的代码字符集的字符个数；

Q_i——第i层（或第i面）的编码对象的总数量。

当为顺序码或系列顺序码时，i=1则以上公式变为：

$$N=\log_a Q \qquad \text{……（2）}$$

9.3　代码类型及形式

代码类型及形式应符合下列要求：

a)代码字符可选择使用数字型代码、字母型代码、字母数字型代码；

b)代码字符应正确无误、易认易读。应避免使用容易被混淆和误解的字符。在一个标准中，音相近、形相似的字符应避免同时出现，如字母“I”与数字“1”；

c)代码最好全部用数字或全部用字母表示。字母数字混合的形式一般在特殊位置使用，不宜在随机的位置使用；

d)采用数字型代码时，如果有收容类目时其代码采用末位数字为“9”的代码；

e)选用顺序码时，代码一般要等长。例如：用001~999，而不用1~999；采用层次码时，同层次的代码要等长；

f)在同一个标准中，代码书写形式要一致，包括字母的大、小写，代码的字体字号。

9.4 代码表格式

数据元值域分类与代码表（或代码表）应以表格的形式列出。依据WS/T 303要求，表格由代码栏（代码指编码值，可简称为“值”）、编码对象名称栏（在代码表中可简称“值含义”）、说明栏组成，并可根据实际需要适当增减栏目。

当表格比较简单时，为了减少篇幅，可以在一页中并排或两列以上的表格。

9.5 代码表书写要求

数据元值域代码表书写要求如下：

a)代码（值）栏：代码一般在代码栏内左起顶格书写；当代码层次较多时，代码栏可按层次再进行划分；

b)名称（值含义）栏：编码对象名称在名称栏内左起顶格书写，每个编码对象名称占一行。当编码对象名称较长时，可延续至下一行，延续部分要与上一行对齐。采用线分类法时，第一层次的编码对象名称左起顶格书写，第二层空一个字，依此类推；

c)说明栏：说明的内容在说明栏左起顶格排。

9.6 代码表命名

代码表应具备在特定领域背景上获得权威认可的名称。代码表的名称应准确反映代码表作为数据元表示类属性之一的特征，不应放大或缩小其使用范围。

代码表的名称应简洁，传达明确的语义，体现代码表的本质内容。

9.7 代码表标识符

代码表应该在特定使用领域内具有唯一的标识符，用来识别表示数据元值域的编码体系。组学数据元值域代码表的标识符根据的归类确定。结构为：

CV+7位数字，组成总长度为11位的字母数字混合码，包括2个分隔符号“.”。

按类别代码、顺序号从左向右顺序排列。结构如图7所示。

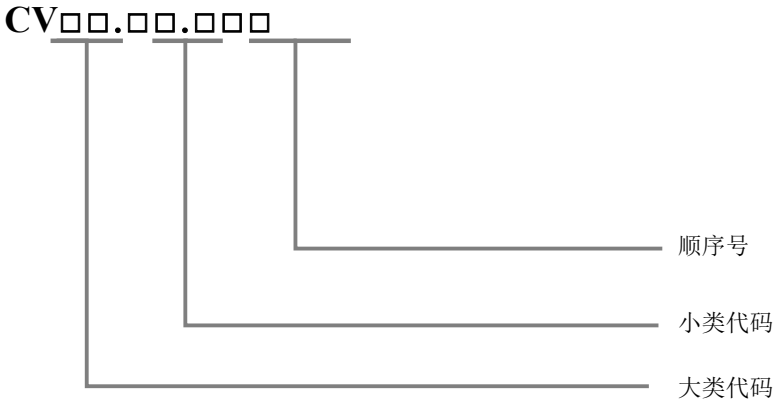

图7 代码表标识符结构

CV——表示数据元值域的编码值，Coded Value；

大类代码——2位数字，表示组学分类中第一层（大类）的代码，见表A.1。

小类代码——2位数字，表示组学分类中第二层（小类）的代码，见表A.2。小类代码和大类代码之间加“.”隔开。

顺序号——用3位数字表示，代表每一类别下值域代码表的序号，数字大小无含义，从001开始编码。顺序号与类别代码之间加“.”隔开。