单细胞组学 宿主单细胞组学数据元数据

版本控制：Version 0.0.7 日期：2024/09/28 (YYYY/MM/DD)

Release Note：完成和现有元数据（metadata）目录对接

目录

[1. 前言/引言（撰写目的 规则 对应的数据资产等级） 3](#_Toc178413685)

[2. 范围 6](#_Toc178413686)

[3. 规范性引用文件 6](#_Toc178413687)

[4. 分类和编码（唯一识符/ID System） 6](#_Toc178413688)

[5. 术语和定义（必备） 7](#_Toc178413689)

[6. 约定 8](#_Toc178413690)

[6.1 概述 8](#_Toc178413691)

[6.2 元数据描述/元数据基本属性 8](#_Toc178413692)

[6.2.1元数据描述标识 8](#_Toc178413693)

[6.2.2元数据术语 9](#_Toc178413694)

[6.2.3元数据含义 9](#_Toc178413695)

[6.2.4元数据限定 9](#_Toc178413696)

[6.2.5元数据场景 9](#_Toc178413697)

[6.3 UML模型图 10](#_Toc178413698)

[6.3.1 UML模型符号 10](#_Toc178413699)

[6.3.2 UML模型关系 11](#_Toc178413700)

[6.3.3 UML模型构造型 13](#_Toc178413701)

[6.3.4 元数据UML模型图 14](#_Toc178413702)

[6.4 数据字典 14](#_Toc178413703)

[7. 要求 14](#_Toc178413704)

[7.1通则 14](#_Toc178413705)

[7.1.1 元数据概念模式 14](#_Toc178413706)

[7.1.2 核心元数据元素 14](#_Toc178413707)

[7.1.3 元数据专用标准的一致性 15](#_Toc178413708)

[~~7.1.4 单细胞组学属性数据的元数据记录~~ 15](#_Toc178413709)

[7.1.5 元数据应用示例 15](#_Toc178413710)

[7.2 元数据包和元数据实体关系 15](#_Toc178413711)

[7.3 基础包说明 16](#_Toc178413712)

[7.3.1 元数据信息(元数据) 16](#_Toc178413713)

[7.3.2 测序技术信息（测序技术） 17](#_Toc178413714)

[7.3.3 样本信息（样本） 17](#_Toc178413715)

[7.3.4 实验设置信息（实验设置） 17](#_Toc178413716)

[7.3.5 质量控制信息（质量控制） 17](#_Toc178413717)

[7.3.6 责任人信息（责任人） 18](#_Toc178413718)

[7.3.7 唯一标识符信息（唯一标识符） 18](#_Toc178413719)

[7.3.8 基因组信息（参考基因组） 18](#_Toc178413720)

[7.3.9 分发信息（分发） 19](#_Toc178413721)

[7.3.10 覆盖范围信息（覆盖范围） 19](#_Toc178413722)

[7.3.11 法律信息（法律） 19](#_Toc178413723)

[7.3.12 实体信息（实体） 19](#_Toc178413724)

[7.3.13 数据产出的受益者信息（受益者） 20](#_Toc178413725)

[8. 元数据模式 20](#_Toc178413726)

[8.1 基础包 20](#_Toc178413727)

[8.1.1 元数据信息 20](#_Toc178413728)

[8.1.2测序技术信息 21](#_Toc178413729)

[8.1.7 唯一标识符信息 22](#_Toc178413730)

[附录A：元数据字典 23](#_Toc178413731)

# 前言/引言（撰写目的 规则 对应的数据资产等级）

起草规则

提出机构

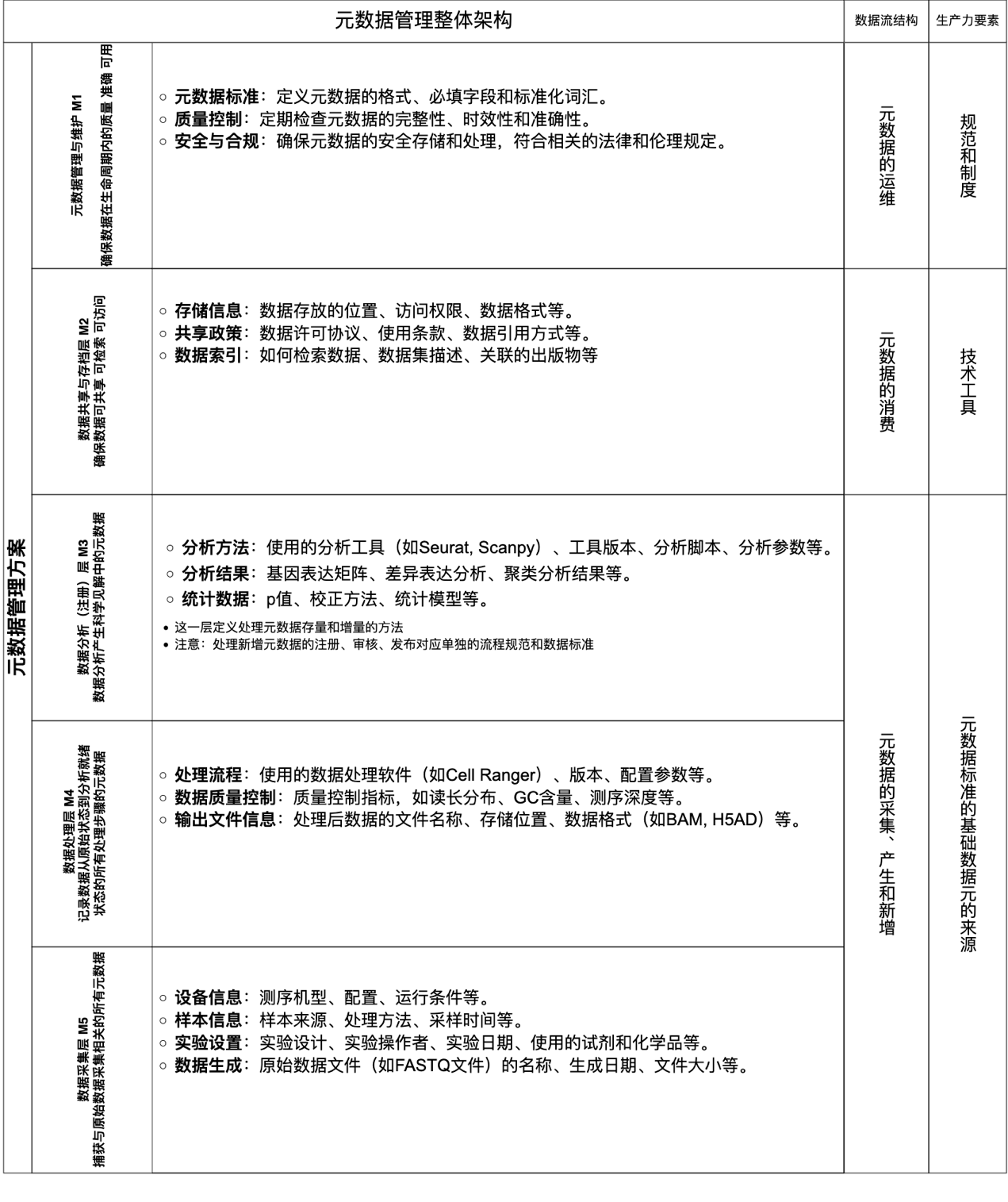
归口单位

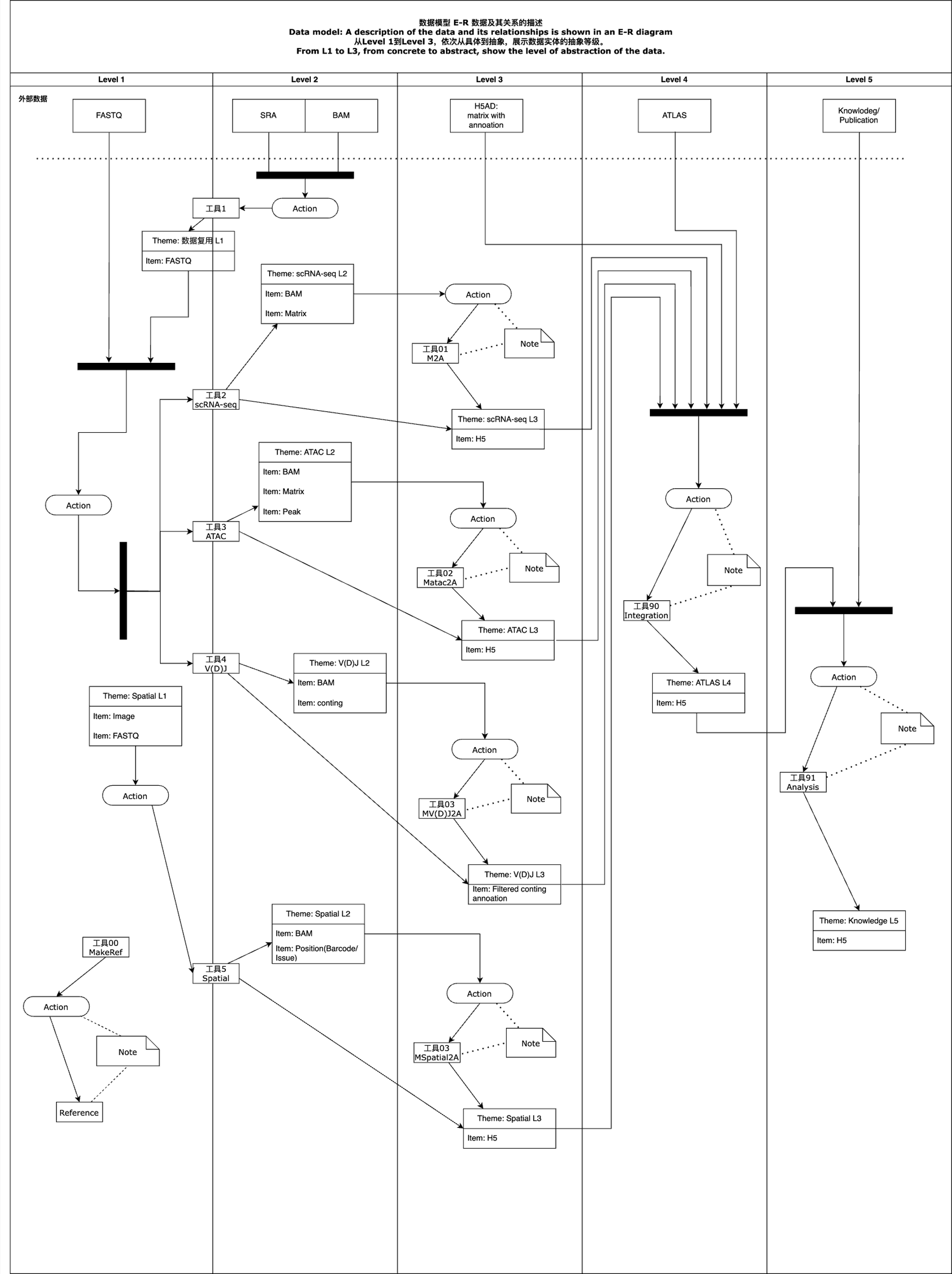
起草单位

起草人

作为数据标准的基本组件之一，本文件为宿主单细胞组学：数据标准系列的基本参照文件，定义了数据标准的最小数据元、数据流模型的基本属性、数据资产实体的全周期检索字段，是比数据标准更为底层的组织共识，也为数据标准的一些列抽象建立坐标系。本文件的撰写顺序参照数据流模型的数据分级进行，其在数据资产架构中属于D2层的元数据与注释管理（Metadata and annotation management）主题域。而在元数据管理架构中，本文件主要服务于元数据的采集、产生和新增。本文件将服务于存量场景中，因此元数据的消费和运维不在此标准的内容中，消费和运维的部分将在其他标准中给出。

本文件是宿主单细胞组学数据集元数据内容的标准。通过提供一个可扩展的、用于描述和归档管理数据的概念模式，规范化元数据的组织管理，促进单细胞组学科学数据的共享和交换，用数据为科研活动赋能。





# 范围

本文件规定了宿主单细胞组学数据主题元数据的要求，技术内容和概念模式。规定了元数据应用必备的最小的数据元素集，以及对元数据进行取舍、扩展和制定元数据标准的专用的规则和方法。

适用于宿主单细胞组学数据集的编目和描述，组织管理以及数据交换网站的数据服务。

# 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。凡是注明日期的应用文件，仅注日期的版本适用于本文件。凡是不住日期的应用文件，其最新版本（包括所有的修改版）适用于本文件

~~GBT 29859-2013 生物信息学术语~~

# 分类和编码（唯一识符/ID System）

用以唯一标识数据元素，属于全局唯一的标准进行构建。

在所有数据标准全部制定后进行定义。

# 术语和定义（必备）

本部分定义了标准中使用的专业术语，确保所有读者对主要内容的理解一致。

内容上，应该至少包括：条目编号、术语、英文对应词、定义。

# 约定

## 6.1 概述

本文件使用元数据基本属性，通过元数据管理架构、UML模型和数据字典两种方式对元数据概念模式进行描述。

## 6.2 元数据描述/元数据基本属性

本文件使用语义和语法两个角度对元数据基本元素和元数据实体进行描述，使用以下几个角度进行定义。

1. 标识：术语唯一标识符、数据类型标识符、数据值域标识符、数据场景标识符；
2. 术语：中文名称、英文对应词、缩写名；
3. 含义：定义；
4. 限定：数据类型、值域；
5. 场景：约束/条件。

### 6.2.1元数据描述标识

1. 术语唯一标识符

元数据元素或元数据实体的中文标记和英文标记及缩写名。元数据实体名称在本文件中是唯一的，元数据元素名称在元数据实体中是唯一的，通过元数据实体名称和元数据元素名称的组合，使元数据元素名称在整个标准中唯一。为了便于在大规模数据库系统中进行索引和使用，对每个元数据术语单位使用唯一标识符号进行标记。

1. 数据类型标识符

元数据元素的数据类型，通过字符串、整数等数据类型对元数据元素进行分类，定义了元数据的有效操作，使元数据元素在系统中实现规范化的存储、使用和新增。

1. 数据值域标识符

元数据元素的数据值域，通过标识限定元数据元素的值域，使元数据元素在存储、使用和新增的过程中的行为可控。

1. 数据场景标识符

元数据元素和元数据实体在存储、使用和新增的过程中有对应的约束/条件。通过对元数据实体和元数据元素的使用场景的标记，其和术语、数据类型、数据值域共同组成对元数据实体和元数据元素的在系统的中的唯一定义，能够通过标识符限定元数据元素和元数据实体的基本信息。

### 6.2.2元数据术语

1. 中文名称

元数据元素或元数据实体的中文标记。元数据实体名称在本文件中是唯一的，元数据元素名称在 元数据实体中是唯一的，通过元数据实体名称和元数据元素名称的组合，使元数据元素名称在整个标准 中唯一。

1. 英文名称

元数据元素或元数据实体的英文标记。

1. 缩写名

元数据元素或元数据实体的英文缩写名称，规则如下：

* 缩写名在本文件范围内应唯一；
* 存在国际或行业领域惯用英文名称缩写的应采用惯用缩写。

### 6.2.3元数据含义

1. 定义

对元数据元素或元数据实体含义的解释，以使之与其他元数据元素或元数据实体在概念上相区别。

### 6.2.4元数据限定

1. 数据类型

元数据元素的数据类型的规定，例如整型、实型、字符串型等。

1. 值域

元数据元素的有效值域和允许对该值域内的值进行有效操作的规定。元数据元素所允许的值的集合。如对于疾病类型元数据，限定为呼吸系统疾病的关键词。

### 6.2.5元数据场景

1. 约束/条件

说明一个元数据实体或元数据元素是否应在元数据中选用，还是有时选用。 该属性可以有如下的值：

* + 必选(M)，元数据实体或元数据元素应选用；
  + 条件必选(C)，说明元数据实体或元素是否选用的条件。当该条件满足时，其实体或元素 成为必选实体或元素；
  + 可选(O)，元数据实体或元数据元素可选用，也可不选用。如一个——可选实体未被选用，则该实体所包含的元素(包括必选元素)也不选用。——可选实体可包含必选元素，但那些元素 只当——可选实体被选用时才成为必选的。

1. 数据唯一性

说明元数据元素或元数据实体是否在场景下具有唯一性，或可以有的实例的最大数目。

## 6.3 UML模型图

### 6.3.1 UML模型符号

本文件采用统一建模语言（UML）描述元数据子集、元数据实体和元数据元素之间的关系，用UML中的包表示元数据子集，类表示元数据实体，属性表示元数据元素。本文件使用的UML符号如图1所示：

图示

描述已自动生成

图1 本文件使用的UML符号

### 6.3.2 UML模型关系

6.3.2.1关联

用于描述两个或更多类之间的一般关系。关联方向可在线段终点用箭头标记，指明方向的关联称 为单向关联。如不指明方向，则假定为双向关联。

6.3.2.2聚合

用于创建两个类之间的部分与整体的关系。在该关系中， 一个类担当容器角色，另一个类担当容器 的构件角色。聚集的表示法是从“部分”类画一条带空心菱形箭头的实线指向“整体”类。

6.3.2.3泛化

表示父类(或超类)与可替代它的子类之间的关系。父类是泛化类，而子类则定义为特化类。泛化 的表示法是从子类画一条带空心三角箭头的实线指向父类。

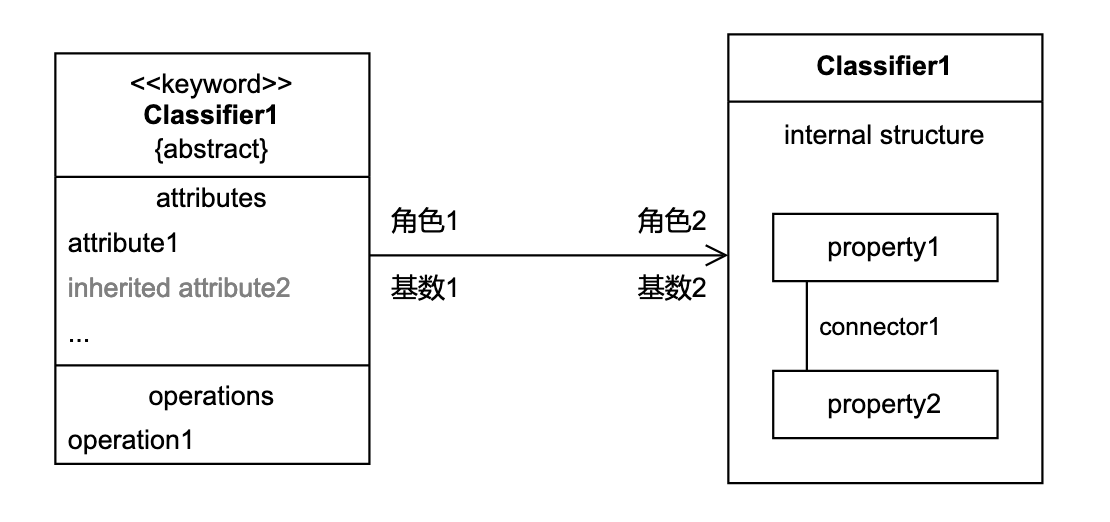
6.3.2.4依赖

用于表示对包(元数据子集)的理解、使用等依赖其他的包(元数据子集)。依赖的表示法是从依赖包画一条带箭头的虚线指向被依赖包。

6.3.2.5角色

表示类在关联中的作用。UML模型中可使用“角色名称”说明目标对象在关联中所承担的角色。在双向关联中，应提供两个类的角色名称，单向关联只需说明已知类的角色名称。图2说明了在UML图中如何表示角色名称和基数。

本标准使用的UML角色含义如图2所示：



注：图2中，角色名称“角色1”是类1对类2的关系，“角色2”是类2对类1的关系，“基数1”表示类1有多少个对 象和类2的单个对象关联，“基数2”表示类2有多少个对象和类1的单个对象关联。

基数的取值可以是一种正整数n(例如“1”“2”“3”等)，也可以是“0..n”（从0到n）、“0..\*”、“n..\*”、“m..n”(m小于n)，还可以是“m,n”(m小于n)。

其中m和n都是一确定的正整数；“\*”代表许多、多个；“..”、“0..×”、“n..\*”语境中表示“或”，例如，基数1为“0..\*”表示类1的0个或多个对象与类2的1个对象关联，基数1为“1..\*”表示类1的1个或多个对象与类2的1个对象关联；“.”在“0..n”、“m..n”语境中表示“到”,例如，基数1为“1..3”表示类1的1个、2个或3个对象与类2的1个对象关联；“,”表示“或”，例如，基数1为“2,4”表示类1的2个或4个对象与类2的1个对象关联。在单向关联中，一般只标注箭头指向的多少个对象与箭头背向的类的一个对象关联。

图2 本文件使用的UML角色的定义

### 6.3.3 UML模型构造型

UML构造型是现有UML概念的扩展机制，用来对其它新增的UML元素进行分类(或标记)的模型元素。构造型在固有的UML元模型类层次结构的基础上，增强了分类（扩展）机制。本文件使用如下构造型：

1. <<类型>>（<<type>>）：一个构造型类，用于定义实例（对象）的域以及作用于这些对象的操作。类型可以有属性和关联，但没有方法。
2. <<数据类型>>（<<dataType>>）：一组（各不相同的）值的描述符，是一个具有很少操作或没有操作的类。它通过值来传递，其操作不改变值，但可以把值作为结果返回。数据类型的主要目的是容纳另一个类的抽象状态，以进行传输、编码或持久地存储。数据类型包括基本预定义 类型(例如，整型、实型、字符串类型等)和用户定义类型。
3. <<值域>>（<<valueRange>>）：值域用于定义属性可以接受的值的范围，通常用于数值型或枚举型属性，以确保数据的有效性和一致性。在定义实验条件、测量值等属性时，可以指定其值域，例如温度范围、浓度范围等。通常值域是封闭的，有明确的规则、描述及所有元素可知。
4. <<枚举>>（<<enumeration>>）：一种数据类型，其实例组成一个有名称的字符值的列表。 枚举名称及其字符值均要声明。枚举意味着一个类中的可能值完全已知。
5. <<代码表>>（<<codeList>>）：用于描述更开放的枚举。它是一种灵活的枚举，用于表示可 能值的一个长列表。如果该列表的元素是完全已知的，应使用枚举；如果仅知道元素的可能值，则应使用代码表。即枚举是封闭的、不可扩展的，而代码表是可扩展的。描述可能值的开放性枚举，适用于值的范围不完全已知的情况。如FASTQ文件的质量值分数的规则有多种，但是均已知，适用于枚举的情况。而测序技术不断进步，具体的技术名称术语则应该使用代码表代表可扩展的枚举。
6. <<抽象>>（<<abstract>>）：抽象类，该类不能被直接实例化，UML模型中用斜体显示这个类的名称。

### 6.3.4 元数据UML模型图

本文件使用UML抽象对象模型定义单细胞组学数据中的元数据的模型。在第8章陈述的UML模型图中提供的UML图展示了由相关元数据实体、元素、数据类型和代码表构成的元数据子集（以UML包的形式表示），即使用uml模型来陈述元数据实体（包）的内部构成。在这些图中，与相关元数据实体（包）的元素有关的内容将被省略，并在元数据实体名称下用圆括号说明其所属的包。6.2规定的五个角度中，反映语法特征的角度包括限定、场景，而UML中类的属性的多重性以及类间关系的基数可看作是元数据元素和元数据实体限定角度的一种体现

## 6.4 数据字典

数据字典对每个元数据元素和元数据实体的中文名称、英文名称、缩写名、定义、数据类型、值域、约束/条件和数据唯一性等对五个角度的描述途径进行了完整说明。在数据字典中，角色名称用来标识元数据抽象模型关联，并由“角色名称：”开头，将它们与其他元数 据元素区分；对于元数据实体的值域，用其包含的行数来说明，使用数据类型属性定义元数据实体、构造型和关联；对于元数据元素和元数据实体的约束/条件属性，用“M” 表示必选，“C”表示条件必选、“O”表示可选；对于元数据元素和元数据实体的最大出现次数，只出现一次用“1”表示，不固定次数的重复出现用“N”表示，允许不为1的固定出现次数，并用相应的数字(即“2”“3”……)表示。

7.要求

# 要求

本部分列出了必须遵循的具体条款和条件，包括元数据的组成、必需元素及一致性标准。

## 7.1通则

本部分提供了制定和实施元数据标准的基本框架，确保元数据描述的完整性和一致性。

### 7.1.1 元数据概念模式

本标准定义描述土壤科学数据所需要的元数据概念模式，元数据描述的对象是“数据集”。元数据实体和元数据元素在第7章、第8章和附录A中阐明。

### 7.1.2 核心元数据元素

元数据记录应包括核心元数据中必选的最少元数据元素，见附录B。

### 7.1.3 元数据专用标准的一致性

元数据专用标准的制定应遵循附录C。其一致性要求如下：

1. 声明与本标准严格一致的元数据专用标准应包含本标准中定义的元数据元素、元数据实体和元数据子集，并能通过附录D中所要求的测试。
2. 声明与本标准一致的元数据专用标准可包含扩展的元数据元素、元数据实体和元数据子集，并能通过附录D中所要求的测试。

### ~~7.1.4 单细胞组学属性数据的元数据记录~~

~~数据集包含单细胞组学属性数据时，元数据记录应包含完整技术名称（参见附录E）。~~

### 7.1.5 元数据应用示例

单细胞组学元数据应用示例参见附录F。

## 7.2 元数据包和元数据实体关系

本文件中，元数据子集用UML包表示。每个包包含一个或多个元数据实体(UML类),元数据实体包含一个或多个元数据元素(UML类属性)。元数据实体可与一个或多个其他元数据实体相关。元数据实体可按需要聚合或重复以满足本文件规定的必选要求以及用户的其他要求。图3表示包的结构，~~其中责任者包和覆盖范围包是构造型实体，其定义见7.3.3~~。第8章和附录A分别用不同的方式描述了元数据。

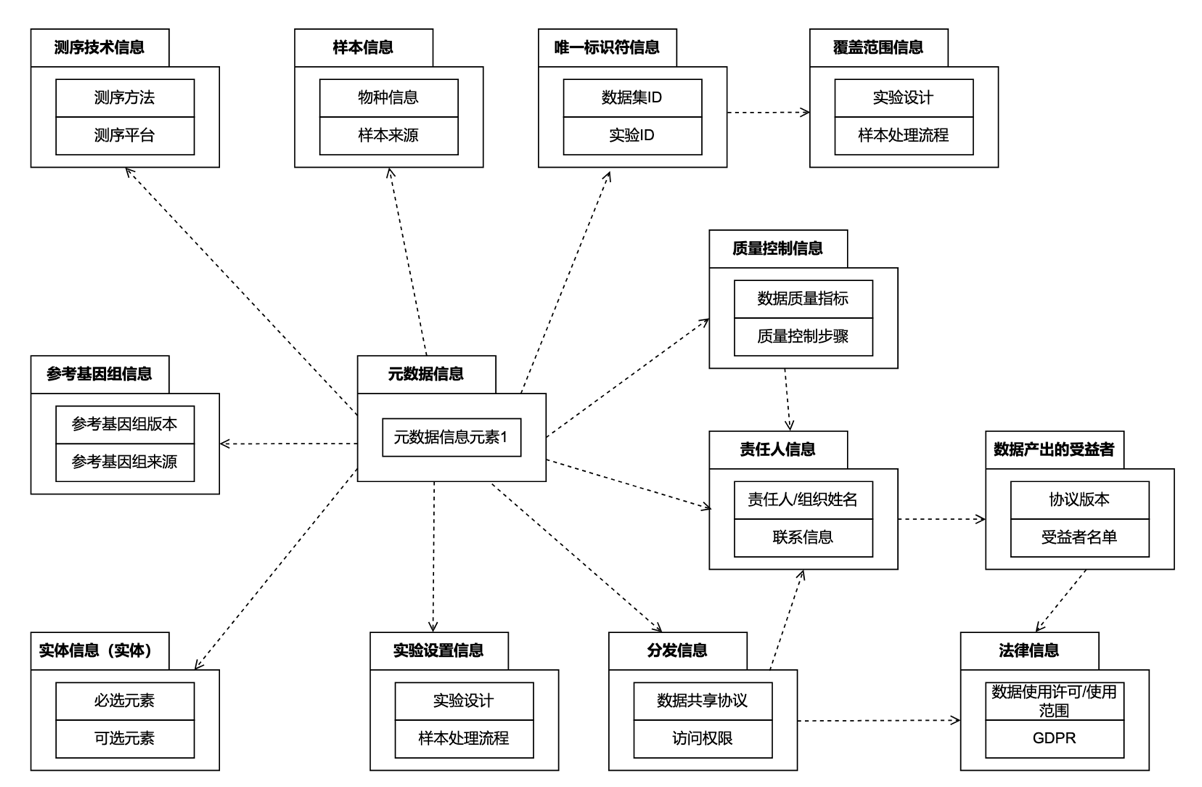


图3 元数据包

表1列出了元数据的包与元数据实体之间的关系，各个包对应的元数据聚集实体列在元数据实体 栏目下。每个包有一对应的条目，包中含有的聚集实体则在7.3、7.4、第8章和附录A中详细说明。

表 1 元数据包和元数据实体间的关系

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 本章条目编号 | 元数据包 | 元数据实体 | UML图 | 数据字典 |
| 7.3.1 | 元数据信息 | 元数据 | 8.1.1 | A.1.1 |
| 7.3.2 | 测序技术信息 | 测序技术 | 8.1.2 | A.1.2 |
| 7.3.3 | 样本信息 | 样本 | 8.1.3 | A.1.3 |
| 7.3.4 | 实验设置信息 | 实验设置 | 8.1.4 | A.1.4 |
| 7.3.5 | 质量控制信息 | 质量控制 | 8.1.5 | A.1.5 |
| 7.3.6 | 责任人信息 | 责任人 | 8.1.6 | A.1.6 |
| 7.3.7 | 唯一标识符信息 | 唯一标识符 | 8.1.7 | A.1.7 |
| 7.3.8 | 参考基因组信息 | 参考基因组 | 8.1.8 | A.1.8 |
| 7.3.9 | 分发信息 | 分发 | 8.1.9 | A.1.9 |
| 7.3.10 | 覆盖范围信息 | 覆盖范围 | 8.1.10 | A.1.10 |
| 7.3.11 | 法律信息 | 法律 | 8.1.11 | A.1.11 |
| 7.3.12 | 实体信息 | 实体 | 8.1.12 | A.1.12 |
| 7.3.13 | 数据产出的受益者信息 | 受益者 | 8.1.13 | A.1.13 |

## 7.3 基础包说明

### 7.3.1 元数据信息(元数据)

“元数据信息”描述单细胞组学科学数据的全部元数据信息，用必选实体“元数据”表示。“元数据”实体由以下内容构成：

——必选元素：元数据创建日期、元数据标准名称和元数据责任者；（元素是元数据的基本组成部分，通常用于描述特定的信息属性。它们是元数据记录中的字段。）

——可选元素：元数据标识、元数据语种、元数据字符集和元数据标准版本；

——必选实体：标识、实验设置、质量控制、测序技术；（实体是更复杂的结构，通常表示一个整体概念或对象，其中包含一个或多个元素。实体可以看作是数据的集合或类别）

——条件必选实体：参考基因组、责任人；（条件必选实体是在特定条件下必须包含的实体。它们并不是在所有情况下都必需，但在某些情况下则是必要的。）

——可选实体：实体、覆盖范围、分发。

### 7.3.2 测序技术信息（测序技术）

测序技术信息描述单细胞测序过程中所使用的技术和平台，必选实体“测序技术”表示。

——必选元素：技术名称、平台类型、测序模式；

——可选元素：技术版本、技术供应商、文献引用；

——可选实体：技术说明、应用实例。

### 7.3.3 样本信息（样本）

样本信息描述单细胞测序分析中使用的样本特征，必选实体“样本”表示。

——必选元素：样本ID、样本来源、样本类型、采集日期；

——可选元素：样本处理流程、样本存储条件；

——可选实体：样本描述、样本组别。

### 7.3.4 实验设置信息（实验设置）

实验设置信息描述进行单细胞测序实验所需的详细设置，必选实体“实验设置”表示。

——必选元素：实验设计、样本处理方法、实验步骤；

——可选元素：实验日期、实验地点、参与人员；

——可选实体：设备信息、实验条件说明；

——条件必选实体：法律信息。

### 7.3.5 质量控制信息（质量控制）

质量控制信息描述对单细胞测序数据进行的质量评估，必选实体“质量控制”表示。

——必选元素：质量控制标准、质量报告、检测结果；

——可选元素：质量控制方法、质量评估工具；

——可选实体：质量审查记录、质量改进措施；

——条件必选实体：责任人信息。

### 7.3.6 责任人信息（责任人）

责任人信息记录与单细胞测序数据相关的责任人的信息，必选实体“责任人”表示。

——必选元素：责任人姓名、联系方式、职务；

——可选元素：责任人所在机构、责任描述、电子邮箱、电话、地址、邮政编码；

——可选实体：协作者信息、联系方式。

### 7.3.7 唯一标识符信息（唯一标识符）

唯一标识符信息定义数据集的唯一标识，使得所有最小元素具备全局唯一性。必选实体“唯一标识符”表示。

——必选元素：数据集ID、DOI、URL；

——可选元素：标识符来源、标识符版本；

——可选实体：标识符历史记录；

——条件必选实体：覆盖范围信息。

### 7.3.8 基因组信息（参考基因组）

参考基因组信息描述用于单细胞测序的参考基因组，条件必选实体“参考基因组”表示。

——必选元素：基因组名称、版本、来源；

——可选元素：基因组注释、基因组下载链接；

——可选实体：基因组比较信息、基因组特征。

### 7.3.9 分发信息（分发）

分发信息描述单细胞测序数据的分发和共享情况，必选实体“分发”表示。

——必选元素：数据共享政策、分发渠道；

——可选元素：分发许可证、分发时间；

——必选实体：数据使用条款；

——条件必选实体：法律信息、责任人信息；

——可选实体：分发统计。

### 7.3.10 覆盖范围信息（覆盖范围）

覆盖范围信息定义数据集所涉及的时间和空间范围，必选实体“覆盖范围”表示。

——必选元素：时间范围、空间范围、技术分类；

——可选元素：地理坐标、时间标记；

——可选实体：相关环境因素、历史数据。

### 7.3.11 法律信息（法律）

法律信息记录与数据集相关的法律条款和要求，必选实体“法律”表示。

——必选元素：法律要求、合规性声明；

——可选元素：法律责任、数据使用协议；

——可选实体：法律文件链接、法律审查记录。

### 7.3.12 实体信息（实体）

实体信息描述数据集的组成部分和结构，必选实体“实体”表示。

——必选元素：实体名或文件名、实体描述；

——可选元素：内部格式、属性概述、其他结构说明；

——可选实体：实体结构图、附加说明。

### 7.3.13 数据产出的受益者信息（受益者）

“数据产出的受益者信息”定义与数据集相关的受益者信息，必选实体“受益者”表示。

——必选元素：受益者姓名、联系方式、受益者角色、受益目的；

——可选元素：受益者单位、受益者简介、受益者协议；

——条件必选实体：责任人、法律信息；

——可选实体：受益者协议文档、受益者声明。

# 元数据模式

## 8.1 基础包

### 8.1.1 元数据信息

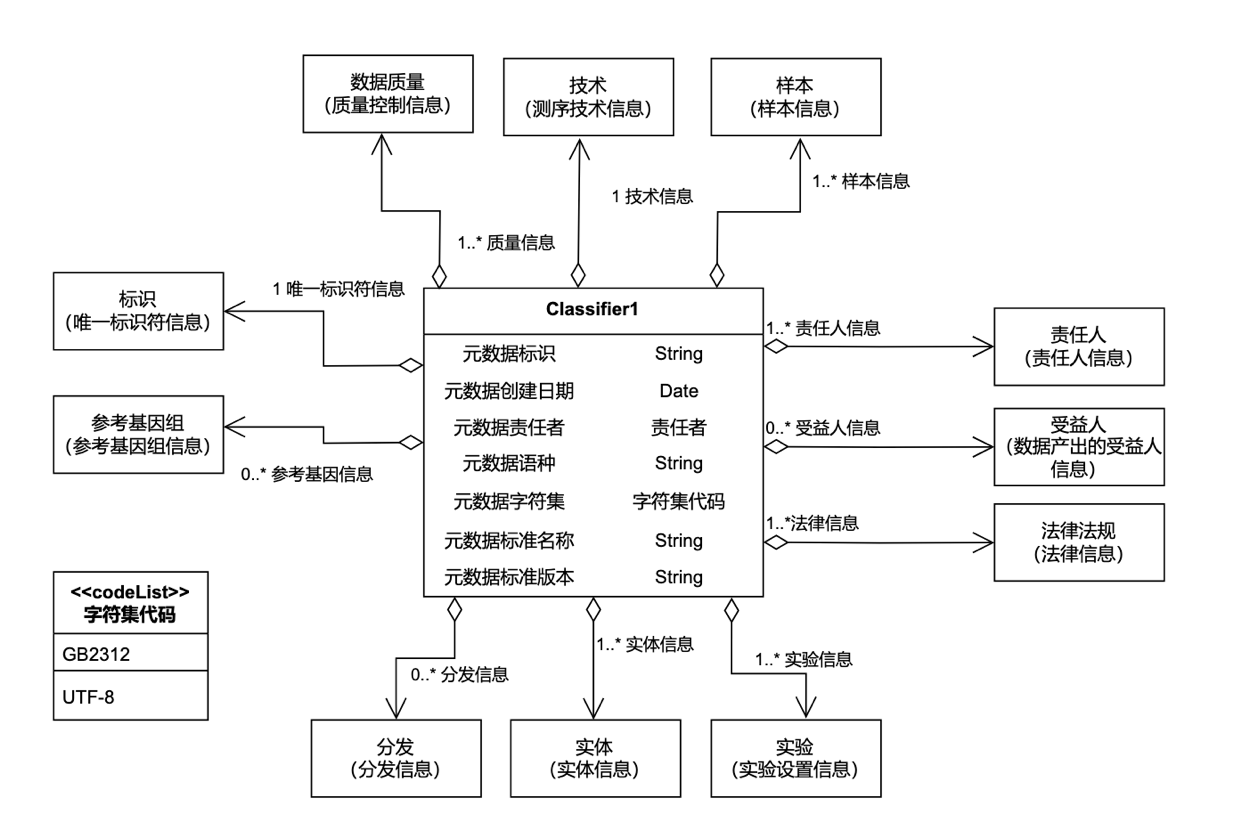
元数据信息见图4。

图4 元数据信息

### 8.1.2测序技术信息

图5定义了单细胞测序数据中测序技术所需的元数据，本图的数据字典见A.1.2

### 8.1.7 唯一标识符信息

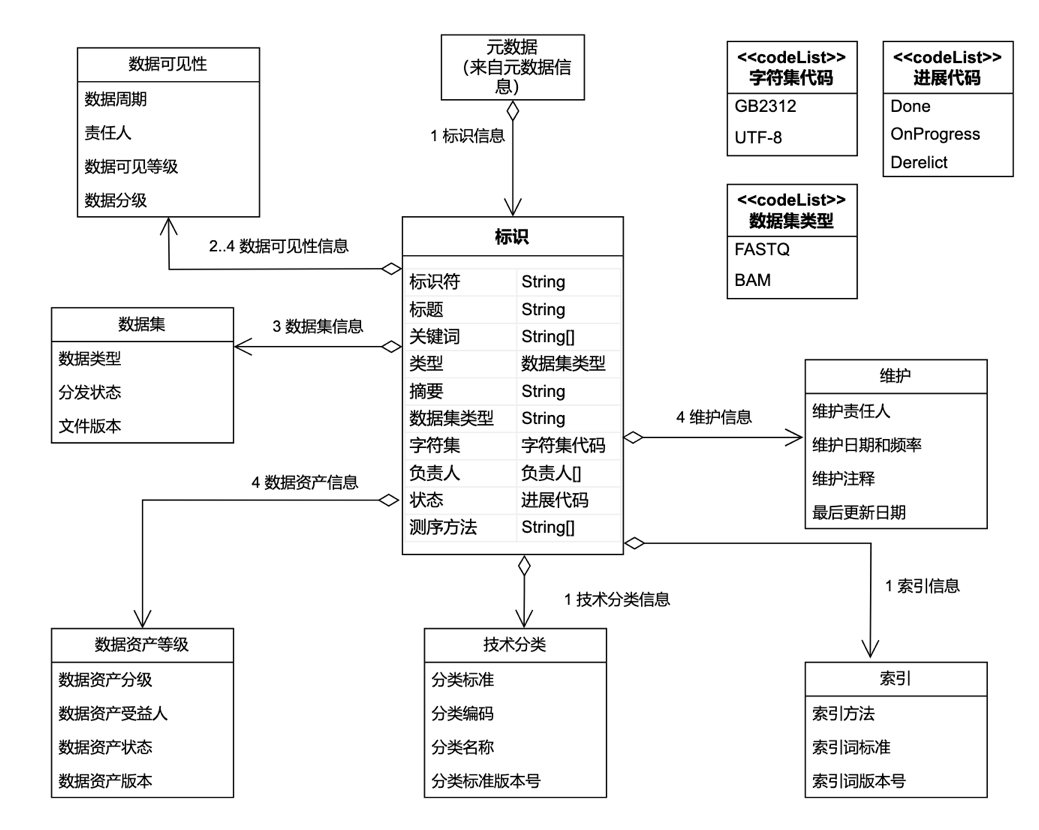
图10定义唯一标识数据集信息所需的元数据。本图的数据字典见A.1.2。

图10 唯一标识符信息

# 附录A：元数据字典

引入现有元数据目录即可