ICS 35.240.70
CCS L 71

团 体 标 准

T/CSTM 00838—2022

材料基因工程 材料数据标识（MID）

Materials genome engineering - Materials data identifier（MID）

2022-08-29 发布　　2022-11-29 实施

中关村材料试验技术联盟　　发布

前　言

本文件参照 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》给出的规则起草。

请注意本文件的某些内容有可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本标准由中国材料与试验团体标准委员会材料基因工程领域委员会（CSTM/FC97）提出。

本标准由中国材料与试验团体标准委员会材料基因工程领域委员会通则技术委员会（CSTM/FC97/TC01）归口。

引　言

材料基因工程是材料科学的新型研发理念。通过从“试错法”向以“数据+人工智能”为标志的数据驱动模式转变，实现新材料及新工艺的理性设计，提高研发效率。数据驱动模式的基础是数据。为了满足材料研究在大数据时代的需求，T/CSTM 00120-2019《材料基因工程数据通则》中定义了“样品信息”、“源数据”和“衍生数据”三类数据，后修订为“样品信息”、“原始数据”和“衍生数据”。以每次操作（样品制备／测量／分析）为条目单位，为每条数据独立赋予唯一且永久的标识符。在数据驱动模式下，基于数字对象标识符和元数据的数字资源注册与管理作为一种有效的技术手段被普遍采用，成为科学数据管理领域成熟的方案。

标识符编码需要遵循唯一性、永久性原则，也要考虑标识符的结构化。标识符的唯一性来自于对编码方式的设计。最简单的唯一性编码是不包含特定结构的数字、字母随机字符串。在标识符中嵌入对应数据的某些特性参数值作为固定字段，如时间、类别、机构代码等，外加顺序号、自定义部分等，可组成具有统一基本结构的标识符。使用者通过具有固定结构的标识符可以快速获取对应数据资源的重要信息，提高使用便捷性。目前，国内外主流的数字标识技术包括数字对象唯一标识符（DOI）、国际标准书号（ISBN）、国际标准连续出版物号(ISSN)、对象标识符（OID）、国际标准关联标识符（ISLI）和科技资源标识（CSTR）等。虽然国内外已有的多种标识符方案均可实现唯一性标识目标，但组成字段的选用往往聚焦于具体领域的具体需求。因此，制定材料基因工程领域的数据标识符是非常有必要的。

材料数据标识的英文表述 Materials Data Identifier，简称 MID，其包含的字段有固定标志代号、产权拥有单位的机构代码、作者在所属单位的个人代码、数据来源代码、注册时间、用户自定义码和系统随机码。海量材料数据的巨大价值只有在材料领域充分实现数据的交换与共享后才能真正体现。MID 中嵌入产权归属单位和数据生产者信息，永久记录数据的知识产权和生产贡献归属，这有助于保障数据所有人的产权利益，完善数据成果的评价激励机制，提高研究者对数据共享的积极性，利于构建良好的数据驱动材料研发生态。另外，以“样品信息”、“原始数据”和“衍生数据”三类数据为参考，在 MID 中嵌入数据来源类别的代码，便于研究者对数据类型辨识，提高目标数据检索效率，更方便地建立用于后续研究所需的数据集。总之，MID 的应用会促进材料大数据的共享，加速数据驱动模式下的材料智能化研发。

材料基因工程 材料数据标识

1 范围

本文件规定了材料数据标识的命名方法和规范。

本文件适用于材料基因工程数据通则框架下的“样品信息”、“原始数据（源数据）”和“衍生数据”的科技资源标识命名规范化，其他类型材料数据的标识符命名规范化可参照执行。

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 2659-2000 世界各国和地区名称代码

T/CSTM 00120 材料基因工程数据通则

3 术语和定义

T/CSTM 00120界定的以及下列术语和定义适用于本文件。

3.1

材料数据标识 materials data identifier; MID

用于唯一标识材料数据资源的一组字符。

3.2

材料数据标识系统 MID system

通过本文件描述的命名方法，以计算机理解的形式，实现对MID进行分配及管理的基础设施。

3.3

材料数据标识命名规则 MID naming convention

MID的构成及字符序列规则，特别是字段、分隔符的构成和字符规则。

3.4

机构代码 organization code

MID系统为全球研究人员所在的单位机构赋予的唯一代码，又称为产权拥有单位的机构代码。法人单位是机构代码分配的最小组织单元。

3.5

个人代码 researcher code

MID系统为各个单位机构的研究人员赋予的唯一代码，又称为作者在所属单位的个人代码。

3.6

数据来源类别 data source categories

材料数据产生的来源类别，分为制备、表征、分析、虚拟制备、虚拟表征五种。

3.7

数据来源代码 data source code

为不同的数据来源类别赋予相应代码，制备、表征、分析、虚拟制备、虚拟表征分别对应S、T、D、M、C。

4 原则

4.1 唯一性

在MID系统中,每个MID仅标识一个数据对象。

4.2 永久性

在对MID进行命名以及在相关的服务或应用中，都不应为MID的存在设定时间限制。当数据对象的所有权、管理责任发生变化时，MID及其数据对象不受影响。

5 MID 命名

5.1 规则

MID由前缀和后缀两部分组成，中间用半角符号“/”分开。前缀包括5个字段，后缀包括2个字段，不同字段之间以半角符号“.”分隔。其他字符采用UTF-8编码。MID命名规则如图1所示。

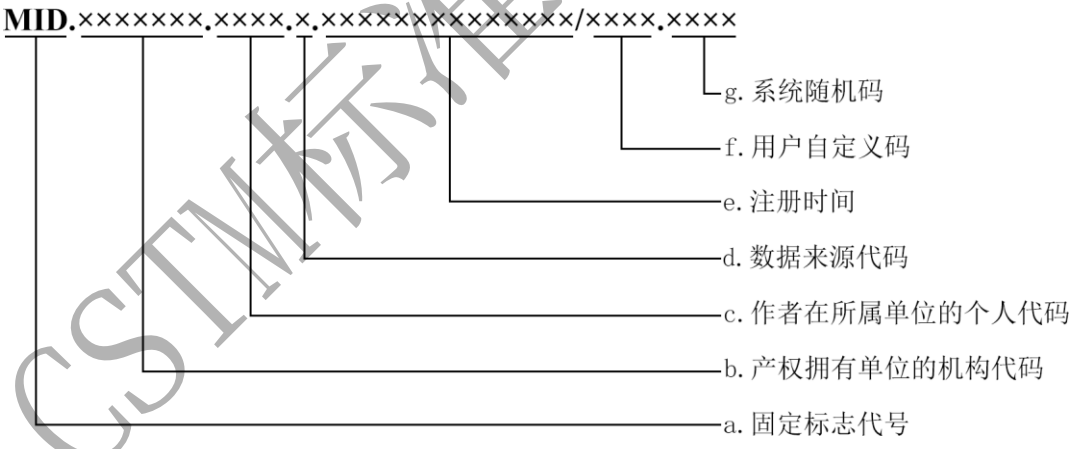

图1 MID命名规则示意图

5.2 前缀

前缀包含a，b，c，d，e共5个字段，具体说明如下：

—— a为标识符的固定标志代号，采用“MID”表示；

—— b为产权拥有单位的机构代码，是由字母和数字构成的7位字符串。其中，前2位字母是国家/地区编号，按GB/T 2659-2000中两字符拉丁字母代码选取。后5位是单位编号，单位类别主要有高等学校、科研机构、企业和其他，单位编号规则见表1。如果单位主体不变，仅变更名称，则编号不变；

表1 单位类别与单位编号对照表

单位类别	单位编号
高等学校（中国大陆）	中国教育部公布的院校招生代码，5 位数字
科研机构（中国大陆）	中国教育部公布的院校招生代码，5 位数字
高等学校（国外和中国香港、中国澳门、中国台湾）	单位名称参考中国教育部教育涉外监管信息网（http://jsj.moe.gov.cn）的公开信息（无编号），单位编号为 5 位数字编号“1xxxx”，数字“1”作为起始字符，后四位编号参考托福送分学校代码（4 位数字）
科研机构（国外和中国香港、中国澳门、中国台湾）	数字“8”作为首位字符，从“80001”开始顺序排号
企业	字母“B”作为首位字符，从“B0001”开始顺序排号
其他	字母“A”作为首位字符，从“A0001”开始顺序排号

—— c为作者在所属单位的个人代码，是由字母、数字构成的4位字符串。这是数据作者在所属单位的代码。同一位数据作者可同时在不同所属单位中拥有不同代码；

—— d为数据来源代码，1位字母，五种数据来源类别分别用五个大写字母表示，具体对应关系见表2；

表2 数据来源类别与代码对照表

数据来源类别	数据来源代码	意义
制备	S	Sample，天然或实验制备的实物样品数据
表征	T	Test，对实物样品实施各种测试产生的表征数据
分析	D	Derived，对表征数据进行分析后产生的结果数据
虚拟制备	M	Model，由计算仿真建模产生的虚拟样品数据
虚拟表征	C	Calculation，对虚拟样品实施计算产生的数据

—— e为注册时间，该字段建议使用年月日时分秒共14位数字表示，用户在使用时亦可根据需求适当扩展字段。该字段代表数据在MID系统注册的时间信息。

例如：20210628201530表示该MID的注册时间为2021年06月28日20时15分30秒。

5.3 后缀

后缀包含f，g共2个字段，具体说明如下：

—— f为用户自定义码，是由字母和数字构成的不定长字符串。数据作者需要针对每一条数据输入相应字符串。用户自定义码可以兼容个人研究过程中对数据的临时命名，最大限度保证研究人员的命名自主权；

—— g为系统随机码，是由字母构成的不定长字符串。该字段由MID系统随机生成，防止MID发生重复。

6 MID 材料数据标识应用示例

MID材料数据标识应用示例见附录A。

附录 A
（资料性）
MID 材料数据标识应用示例

A.1 示例 1：中国研究人员申请制备数据 MID

以下为某组合材料芯片 XRD 表征数据申请标识符的示例。

上海交通大学的研究人员李某某对 Fe-Co-Ni 组合材料芯片逐点进行 X 射线衍射表征，将 XRD 表征数据根据数据库的输入要求进行标准化，对应条目中分别记录对应的实验时间、地点、人员、实验方法、实验装置、实验条件、表征点的几何坐标信息和仪器输出的原始数据（源数据），然后对每一次 XRD 表征数据赋予一个独立的数据标识。表 A.1 展示了其中一个点的 XRD 表征数据申请 MID 时输入的元数据信息。

上海交通大学单位编号 CN10248，李某某个人编号 0009，标识符申请时间 2022 年 7 月 1 日 10 时 25 分 20 秒。标识符系统分配的数据标识符为：MID.CN10248.0009.T.20220701102520/v0006.BFCD。

表 A.1 某组合材料芯片 XRD 数据注册 MID 的元数据表单

序号	项目	子项目	示例或选项
1	数据标题		Fe-Co-Ni 组合薄膜的 XRD 表征数据
2	作者信息	作者姓名	李某某
		作者单位	上海交通大学
3	摘要		采用高通量离子束溅射系统制备了 Fe-Co-Ni 组合薄膜。系统地研究了作为涂层顺序和调制周期的函数进行热处理的此类薄膜的相演变和非晶稳定性。成分结构图是通过自动处理高通量 X 射线衍射获得的数据构建的。
4	来源类别		表征
5	数据 URL		http://www.***.***com/10.10072F00394-012-030-8
6	用户自定义码		v0006
7	关联 MID		MID.CN10248.0009.S.20220601102356/0021.SFAQ

A.2 示例 2：中国研究人员申请虚拟表征数据 MID

以下为某合金材料样品高通量计算数据申请标识符的示例。

清华大学的研究人员张某某获得了基于高通量第一性原理计算的高温合金共格增强相结构稳定性数据。以所构造的每一种 $A_xB_yC_z$（A、B、C 指化学元素，x、y、z 指各元素的分数）化合物为一个样品，例如 Co, Co_3Si, Co_3Al, $Co_3Si_{0.5}W_{0.5}$ 等。在样品信息库对应条目中记录其化学信息、构造信息、结构信息等。对每一个样品利用第一性原理计算软件分别计算总能量、弹性常数、结构参数等。将计算数据根据数据库的要求进行标准化录入，在数据库对应条目中记录地点、人员、计算资源、计算软件、版本号、使用模块和计算条件等参数，以及计算所产生的输出数据等。然后，对每一个样品的计算数据赋予独立的数据标识。表 A.2 展示了其中一个样品的计算数据（虚拟表征）申请 MID 时输入的元数据信息。

清华大学单位编号 CN10003，张某某个人编号 2025，标识符申请时间 2022 年 8 月 23 日 09 时 12 分 36 秒。标识符系统分配的 MID 标识为：MID.CN10003.2025.C.20220823091236/0508.DFCR。

表 A.2 某高温合金共格增强相结构稳定性数据注册 MID 的元数据表单

序号	项目	子项目	示例或选项
1	数据标题		基于高通量第一性原理计算的高温合金共格增强相结构稳定性数据

2	作者信息	作者姓名	张某某
		作者单位	清华大学
3	摘要		本数据为基于高通量第一性原理计算的高温合金共格增强相结构稳定性数据。以所构造的每一种 $A_xB_yC_z$（A、B、C 指化学元素，x、y、z 指各元素的分数）化合物为一个样品，例如 Co, Co_3Si, Co_3Al, $Co_3Si_{0.5}W_{0.5}$ 等。在样品信息库对应条目中记录其化学信息、构造信息、结构信息等。对每一个样品利用第一性原理计算软件分别计算总能量、弹性常数、结构参数等。
4	来源类别		虚拟表征
5	数据 URL		http://www.***.***com/article/pii/030406819300515
6	用户自定义码		0508
7	关联 MID		MID.CN10003.2025.C.20210821091203/1121.AFXZ

A.3 示例 3：美国研究人员申请表征数据 MID

以下为$(Nd_{1-x}Ce_x)_2Fe_{14-y}Co_yB$ 块状材料的磁性表征数据申请标识符的示例。

美国爱荷华州立大学（Iowa State University）研究人员 David 对$(Nd_{1-x}Ce_x)_2Fe_{14-y}Co_yB$ 化合物的磁性进行表征，将磁性表征数据根据数据库的录入要求进行标准化，对应条目中记录对应的实验时间、地点、人员、实验方法、实验装置、实验条件和仪器输出的原始数据（源数据），然后对每一次材料磁性表征数据赋予一个独立的数据标识。表 A.3 展示了一条表征数据申请标识符时输入的元数据信息。

爱荷华州立大学单位编号 US16306，David 个人编号 0315，标识符申请时间 2021 年 10 月 11 日 16 时 37 分 55 秒。标识符系统分配的标识符为：MID.US16306.0315.T.20211011163755/S3553.DEAX。

表 A.3 某$(Nd_{1-x}Ce_x)2Fe_{14-y}Co_yB$ 块状材料的磁性表征数据注册标识符的元数据表单

序号	项目	子项目	示例或选项
1	数据标题		$(Nd_{1-x}Ce_x)_2Fe_{14-y}Co_yB$ 块状材料的磁性表征数据
2	作者信息	作者姓名	David
		作者单位	爱荷华州立大学（Iowa State University）
3	摘要		Ce、Co 共掺杂 $(Nd_{1-x}Ce_x)_2Fe_{14-y}Co_yB$ 化合物的磁性能已在块状多晶和快速凝固的纳米带中进行了研究。在某些特定的 Ce 掺杂浓度下，材料表现出低于 140 K 的自旋重定向转变。与永磁体的应用相关的居里温度、饱和磁化强度和其他磁特性均与 Ce、Co 掺杂有明显关联。
4	来源类别		表征
5	数据 URL		http://www.***.***com/retrieve/pii/S0023643812001934
6	用户自定义码		S3553
7	关联标识符		MID.US16306.0315.S.20211001023302/2323.AQAV

附录 B
（资料性）
起草单位和主要起草人

本文件起草单位：上海交通大学、钢铁研究总院有限公司、北京科技大学、中国建材检验认证集团股份有限公司、中关村材料试验技术联盟、成都材智科技有限公司、中国科学院硅酸盐研究所、西北工业大学、昆明贵研新材料科技有限公司。

本文件主要起草人：汪洪、张澜庭、蔡永珠、路勇超、余宁、唐凌天、沈学静、陈永彦、孙璧瑶、黄海友、包亦望、王蓬、王卓、刘茜、李金山、王毅、张爱敏、陈力。

参 考 文 献

[1] GB/T 18391.1-2009 信息技术 元数据注册系统(MDR) 第 1 部分：框架

[2] GB/T 26816-2011 信息资源核心元数据

[3] GB/T 32843-2016 科技资源标识

[4] GB/T 26231-2017 信息技术 开放系统互连 对象标识符（OID）的国家编号体系和操作规程

[5] GB/T 36369-2018 信息与文献 数字对象唯一标识符系统

[6] ISO2108 Information and documentation—International standard book number (ISBN).

[7] ISO3297 Information and documentation—International standard serial number (ISSN).

[8] ISO17316:2015 Information and documentation — International standard link identifier (ISLI).

[9] Hey T, Tansley S, Tolle KM, editors. The fourth paradigm: data-intensive scientific discovery. Redmond: Microsoft Research Press; 2009.

[10] 汪洪, 项晓东, 张澜庭. 数据+人工智能是材料基因工程的核心[J]. 科技导报, 2018, 36(14): 15-21.

[11] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3(1):160018.

[12] CNRI. Handle System [OL]. http://www.handle.net/.