ICS 01.120
A 00

团 体 标 准

T/CSTM 00120-2019

材料基因工程数据通则

General rule for materials genome engineering data

2019-08-13 发布　　2019-11-13 实施

中关村材料试验技术联盟　　发布

前　言

本标准按照 GB/T1.1—2009 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)通则委员会提出。

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（CSTM/FC97)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的转变，实现新材料及工艺的理性设计。在此模式下，材料研究活动围绕数据产生与数据处理展开，使掌握成分-组织-工艺-性能间关联规律的速度更快、效率更高、成本更少。它代表了材料基因工程的核心理念与发展方向。材料基因组（Materials Genome）这个名词的出现有感于人类基因组计划的成功，但迄今为止并无特定的科学定义，目前的共识是将材料基因工程作为设计预测型材料研发模式的代称。

材料基因工程数据库是实施数据驱动材料科学的基础条件之一，需要收录符合 FAIR（Findable, Accessible, Interoperable, Reusable，可发现、可获取、可互操作、可再利用）原则的数据资源，供社会共享。其中“可发现”指数据及其元数据被赋予全球性唯一并持久的标识，数据被丰富的元数据所描述并在可检索的源中登记或建立索引，易于被第三方(人员与机器)方便地找到；“可获取”指数据及其元数据可使用标准通讯协议通过标识查询并获取；“可互操作”指数据及其元数据的表达使用正式、可获得、共享和广泛使用的语言；“可再利用”指数据及其元数据由多种准确并相关的特征所描述，与细致的出处信息相关联并符合相关领域的标准，从而被不同用户(人员与机器)方便地使用。

数据，特别是源数据(即由测量或计算获得的未经进一步分析的数据)的可再利用性是材料基因工程的重要特征。以某一合金的 X 射线衍射图为例，它可用于获得材料的晶体结构，也可用于分析结晶程度、晶粒大小、晶体取向等参数，还可用于分析合金的相组成。因此一组源数据在不同的使用者手中可以根据各自关切产出不同的结果。传统材料数据库一般仅收集由源数据处理而得到的分析结果（如各种材料性能参数等），而源数据本身通常分散在实验者手中，不被收录。同时，与数据相关的元数据通常也不在收录之列，难以满足 FAIR 原则。因此，有必要建立一种适合材料基因工程需求的数据标准，规范数据的产生过程中必须收集的信息与遵循的格式，以确保数据满足 FAIR 原则，从而得到充分有效的利用。

本通则应对材料科学在数据驱动模式下对数据的需求，将数据分为样品信息、源数据（未经处理的数据）与衍生数据（经分析处理得到的数据）三类，以操作（样品制备/表征/计算/数据处理）为条目单位，对每次操作分别赋予独立资源标识（根据国标 GB/T 32843 或 DOI）。每条数据收集与操作相关的元数据，为样品与数据重复利用提供必要条件。这里，样品可以是实验产生的实物，也可以是经计算产生的虚拟物。同理，原始数据可以来自于表征或是直接的测量，也可以通过模拟计算产生。

为了收录足够元数据与原始数据，本通则兼顾了材料数据专用性与通用性。单个数据条目拥有独立的科技资源标识，独立存在，保证专用性。在使用中，每次分析使用的数据或数据集通过规范化的标准词汇表进行检索，随时建立，保证通用性。同时，对于特定的制备、表征、计算流程也需要建立数据与元数据的格式标准，简化数据存储、分析中的资源消耗，特别是繁琐的元数据录入可以通过高通量制备、表征、计算在操作过程中由计算机自动生成，这对于收集与处理批量产生的数据是必不可少的。

材料基因工程数据通则

1 范围

本通则规定了材料基因工程数据库中收录的数据的内容，即数据的构成及其中必须包含的信息。

本通则适用于所有材料基因工程数据库及其中收录的数据，包括材料样品、原始数据及经过分析处理得出的结果数据。

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，仅所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/T 32843 科技资源标识

3 术语和定义

下列术语和定义适用于本文件。

3.1

实际样品 Actual Specimen

天然或实验产生的实物材料。

3.2

虚拟样品 Virtual Specimen

由计算仿真产生的虚拟材料。

3.3

元数据 Meta Data

与样品和数据的条件有关的数据。

3.4

源数据 Source Data

测量或计算产生的原始数据。

3.5

衍生数据 Derived Data

对源数据或者衍生数据进行分析后产生的结果数据。

4 数据结构

4.1 数据结构基本框架

4.1.1 按材料科学在数据驱动模式下对数据的需求，材料基因工程数据库框架见图1。

4.1.2　按GB/T 32843规定或DOI识别码，以操作为条目单位，对每次操作分别赋予独立资源标识。每条数据收集与操作相关的元数据，为样品与数据重复利用提供必要条件。

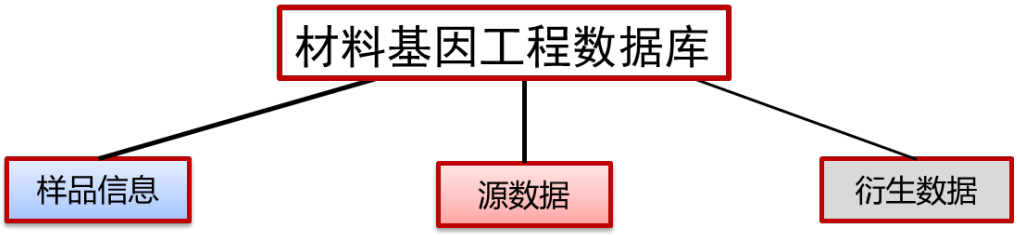

图 1

5　数据内容

5.1　样品信息内容

5.1.1　样品信息内容结构

样品信息内容结构见图2。

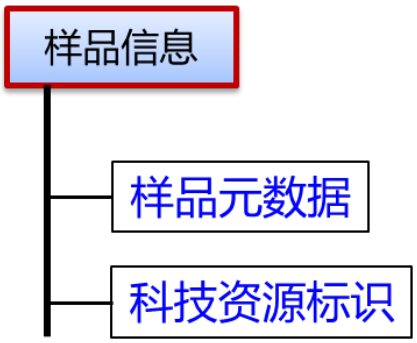

图 2

5.1.2　标准样品内容类型

5.1.2.1　实际样品

材料实际样品的信息包括：

1）规格和名称；

2）产生该实际样品的制备方法和制备条件等元数据；

3）独特且持久标识、如按根据GB/T 32843生成的样品资源标识码或DOI识别码等。

5.1.2.2　虚拟样品

材料虚拟样品的信息包括：

1）规格和名称；

2）产生该虚拟样品的计算方法和条件等元数据；

3）独特且持久标识、如按GB/T 32843生成的样品资源标识码或DOI识别码等。

5.2 源数据内容

5.2.1 源数据结构

源数据结构见图3。

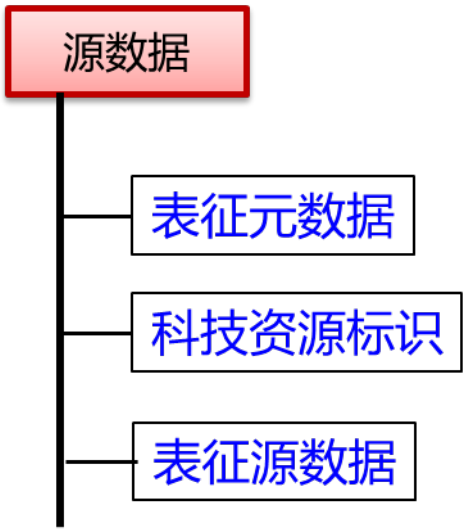

图 3

5.2.2 源数据内容类型

5.2.2.1 实际样品

材料实际样品的源数据信息包括：

1）该次表征、测试实验的方法、条件、样品标识码；

2）该次表征、测试实验的未经处理的数据；

3）该次表征、测试实验的独特且持久的标识、如按GB/T 32843生成的样品表征数据资源标识码或DOI识别码等。

5.2.2.2. 虚拟样品

材料虚拟样品的源数据信息包括：

1）该次计算实验的元数据，如方法和条件、虚拟样品的标识码；

2）计算实验的未经处理的数据；

3）该次计算实验的独特标识、如按GB/T 32843生成的样品表征数据资源标识码或DOI识别码等。

5.3 衍生数据内容

5.3.1 衍生数据结构

衍生数据结构见图4。

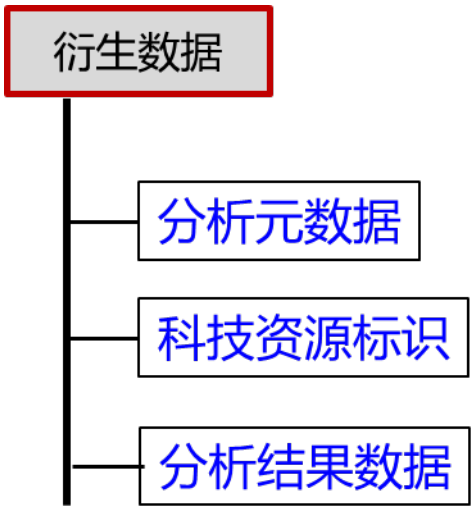

图 4

5.3.2 衍生数据内容来源

5.3.2.1 材料样品的衍生数据内容来源：

1）对源数据或已经分析处理的数据进行分析处理。包括分析元数据，如该次分析涉及的源数据（组）、该次分析处理使用的分析方法和条件等；

2）该次分析处理后得到的分析结果数据；

3）还包括该次分析处理的独特且持久的标识、如按GB/T 32843生成的样品表征数据资源标识码或DOI识别码等。

附 录 A

（资料性附录）

本标准起草单位： 上海交通大学、四川大学、北京科技大学、中国科学院上海硅酸盐研究所、南方科技大学、钢研纳克检测技术股份有限公司、国标（北京）检验认证有限公司、成都材智科技有限公司、西北工业大学、中国工程物理研究院材料研究所、湖南大学、国检集团、中国航发北京航空材料研究院、烟台大学、清华大学、北京航空航天大学、中南大学、中国科学院计算机网络信息中心、中国石化上海石油化工研究院、华南理工大学、北京应用物理与计算数学研究所、上海大学、中国科学院物理研究所、中国科学院金属研究所、中国科学院北京综合研究中心、宁波星河材料科技有限公司、重庆大学、南京工业大学、北京大学、贵研铂业股份有限公司、中国科学院高能物理研究所、中国航空综合技术研究所、苏州热工研究院有限公司、上海华谊集团股份有限公司。

本标准主要起草人：汪洪、张澜庭、杨明理、宿彦京、刘茜、项晓东、贾云海、马通达、王卓、王毅、尹海清、尹嘉清、田泽安、包亦望、刘昌奎、江亮、许庆彦、孙志梅、孙淮、杜勇、李大永、杨小渝、杨为民、杨中民、余宁、宋海峰、张晓彤、张博锋、范晓丽、金魁、周科朝、赵雷、施思齐、班晓娟、顾剑锋、钱权、徐东生、高琛、郭鸿杰、黄晓旭、崔予文、董超芳、鲁晓刚、曾小勤、谢明、鲍华、黎刚、滕春禹、潘峰、薛飞、魏建华。