

科技评论

面向人工智能的材料数据资源基础设施与数据标准化

汪洪^{1,2,3,4}

摘要 人工智能(artificial intelligence, AI)正将材料研发推向“数据驱动”新范式,数据是 AI 赋能材料(AI for Materials, AI4M)成功的关键基础,当前的最大瓶颈是缺乏“机器随手即用”的高质量数据集。2025 年发布的美国“创世使命”(Genesis Mission)专项行动和英国《人工智能赋能科学战略》(《AI for Science Strategy》)均把协同运行的“高质量科研数据集+超级算力”列为国家级战略基础设施。中国《新材料大数据中心总体建设方案》亦明确提出整合全国力量,构建全球最丰富的材料数据基础设施的发展战略,以此支撑 AI 赋能的原始创新与产业应用。然而,数据集只是更大的信息整体的一个子集,是构建者科学认识的反映,带有天然的局限性。当前有效的数据集,未来不一定适用。系统分析了 AI 对材料数据内容与形式的新要求,提出面向 AI 时代的材料数据基础设施,应以全信息数据池为根、高质量数据集/语料库为叶,构建可持续演进的“数据树”。通过对根资源前置标准化,实现原始数据的多次复用与多用途复用,以满足专用模型与通用大模型不断发展的需求。设计了一种模块化可组装数据模型,提出了相应的标准化原则与路径,以解决由材料数据参量的多样性与多变性带来的标准化难题。建设国家级材料大数据资源基础设施是一项长期性的事业,关系到中国新材料领域的长远发展。须做好前瞻性的规划,既立足当下,做好数据集,解决眼前的急需,更要放眼未来,做好数据源,为未来的发展留出足够空间。因此,具有公益与公共属性的材料数据基础设施应具有“池仓一体”的构架,兼顾数据集与全信息数据池,平衡当下与长远。

关键词 数据基础设施;数据集;数据源;数据标准化;全信息数据

随着大数据、人工智能
(artificial intelligence, AI)等数据

1. 上海交通大学材料科学与工程学院, 上海 200240
2. 苏州国家实验室, 苏州 215123
3. 上海交通大学材料基因组联合研究中心, 上海 200240
4. 上海交通大学张江高等研究院, 上海 201203

技术的快速发展,材料科学领域迎来一场深刻变革。以 AI 为内核的数据驱动研发新范式,正成为超越传统的新质力量。数据是 AI 赋能材料(AI for Materials, AI4M)研究的关键要素,数据+AI 共同构成了数据驱动范式的核心内容^[1]。当前一个日益增长的全球性共识是,21 世纪的科学领导地位将取决于谁能最有效地将数据、算力和算

法融合为一个持续循环的学习、实验与创新体系,开发出科学领域的基础大模型。因此,中国、美国、英国等科技大国在最近 2 年动作频频,在国家层面进行了战略部署与资源投入,以求在全球科技竞争中夺得先机。

美国政府于 2025 年 11 月 24 日宣布启动一项由政府主导的名为“创世使命”(Genesis Mission)

收稿日期: 2025-12-18; 修回日期: 2026-02-27

基金项目: 苏州国家实验室自立项目(SK-1702-2025-079)

作者简介: 汪洪,教授,研究方向为人工智能在材料中的应用、材料数据资源与标准化,电子邮箱: hongwang2@sjtu.edu.cn

引用格式: 汪洪. 面向人工智能的材料数据资源基础设施与数据标准化[J]. 科技导报, 2026, 44(15): 25-33; doi:10.3981/j.issn.1000-7857.2025.12.00090

的专项行动^[2],旨在围绕 AI 重建国家的科学引擎,将美国最强大的超级计算机、最具价值的科学数据集和最先进的 AI 系统整合为一个统一平台,成为国家级的战略基础设施,以此加速能源、生物技术、材料和国家安全等领域的科学发现。这项计划不再将 AI 视为传统研究的附加工具,而是将其直接定位为未来突破的核心机制。白宫称这次行动的紧迫性与雄心堪比曼哈顿计划,意在通过对科研数据的整合,最终形成科技版 GPT。按此趋势,在未来 10 年中,全球竞争的焦点将不再是消费 AI,而是科学 AI。

与此同时,英国科学、创新与技术部也发布了《人工智能赋能科学战略》(《AI for Science Strategy》)^[3],提出以数据、算力、人才与文化 3 大支柱和 AI 驱动科学为核心抓手,面向工程生物、聚变能源、先进材料、医学研究和量子技术 5 大优先领域,打造高质量科研数据集和与国家超算平台协同运行的数据与算力基础设施,扩大超算资源开放,培育跨学科科研与技术人才,推动自主实验室、AI 科学智能体等前沿工具在重点领域布局应用,重塑科研组织和发现范式。

中国近年来高度重视 AI 发展,欲以 AI 引领科研范式变革,全面加速各领域科技创新突破^[4],其中特别强调 AI 和产业的融合,为高质量发展提供新动能。在材料领域,国家科技部门自“十三五”开始对材料基因工程领域进行了持续支持^[5-6]。在系统性建立高通量计算和高通量实验基础平台的同时,也不断认识到数据基础设施

的重要性^[7-10]。工业和信息化部、财政部、国家数据局于 2024 年 10 月联合印发《新材料大数据中心总体建设方案》^[11],明确提出集全国之力,整合海量的材料数据资源,搭建中国新材料大数据中心架构体系的目标,用以推动材料数据的汇聚、处理和开发,充分发挥大数据、AI 对新材料产业的技术支撑作用,加速材料研发。

发挥 AI 的效能需要解决数据够用与好用的问题。材料数据具有实践特征,非刻意而不得,因此存在绝对数据量“少”的先天局限。Wang 等^[1,10]提出了“数据工厂”设想,通过建设专用数据生产设施,以流水线般系统性地批量生产、处理标准化的材料数据,解决长期以来数量不足的问题,相关工作已在部署之中。所谓数据好用,就是数据的内容与形式要与 AI 的需求相匹配(AI-ready),使得其信息得以充分挖掘,技术潜力能够充分发挥。当前 AI4M 的核心瓶颈在于缺乏可直接调用、标准化、可共享、带标签/特征说明等特性的高质量数据集^[12]。没有高质量数据集的支撑,再先进的算法也只能“巧妇难为无米之炊”。从目前的实际情况看,国际上各主要材料数据基础设施^[13-15]与中国建设中的新材料大数据中心各数据资源节点均着力于回应当下的需要,以提供相关专业领域所需的高质量数据集为主要目标。

然而,数据集是基于特定目标,选取参数并经处理构建而成的数据集合。受到体量与认识角度的局限,不可能包罗万象。AI 日新月异,新能力、新应用层出不

穷,对数据的内容与形式的要求也不断发展。满足当前的数据资源,未必适应未来的变化。因此,数据资源基础设施的设计须具备前瞻性思维,在动态变化中锚定“信息”这一不变的本质,构建持续演进、长期可用的材料数据资源体系。

材料数据的标准化直接决定了 AI 技术在材料领域的应用深度与成效,是构建数据资源基础设施的基础技术。然而,中国材料数据领域仍面临质量标准不统一、共享机制不健全、AI 适配性不足等突出短板,制约了新材料创新效能的全面提升。AI 时代材料数据标准化的目的不仅仅是达到数据形式格式的统一,更重要的是规范数据内容,以保证信息的完整性,强化 AI 的效果。因此,数据标准需以 AI 应用为导向,根据数据资源类型进行有针对性的顶层设计,从而有效服务于近期与长远的需求。

本文系统分析了 AI 时代对材料数据内容与形式的要求,提出面向未来的数据资源基础设施(如新材料大数据中心资源节点或各种专业数据库),不仅要提供服务相关领域的高质量数据集,还应在数据源层面打造成材料样品全信息数据的集散中心,从而满足 AI 对高质量数据资源不断挖掘的需求。为此,本文设计了一种模块化可组装数据结构,以解决由全信息材料数据参量的多样性与多变性带来的标准化难题,并提出相应的标准化方法与原则。

1 AI 需要什么样的数据

简单说来,数据就是通过观

察、实验或计算得出的结果^[16]。材料是一个极其复杂的系统,材料的性能反映了在其成分、制备工艺等多重因素作用下产生的特定原子组装构造形式在特定条件下的表现。因此,材料数据是这些因素综合效果的信息载体。在传统认知中,数据的核心作用是提供事实,作为科学研究、技术设计、查证及决策所依托的数值支撑。在AI背景下,数据的作用发生了质的转变,为训练、验证和测试AI模型提供信息单元,成为数据挖掘的信息源泉与模型学习、决策的基础。鉴于AI在数据使用中表现出的新特点,对数据在组织形式和内容上都提出了不同于以往的新要求。符合AI要求的数据称为AI适配(即AI-ready)的数据。针对不同场景,AI-ready的含义也有所不同。

表1总结了不同场景下,材料数据的用途、目的、作用、形式与特点。传统模式下,数据的主要作用是作为材料选择的依据,或是在解析形式的材料计算模型或公式中代入的参数。因此,它以单条数据形式出现,每次使用一条,其内容反映相关材料的某种特性。在数据驱动模式下,作为高维度、强耦合、非线性复杂体系中信息元素的承载者,这时数据的使用形式是多模态高维度数据集。数据集

是为了某一特定分析、建模或应用,按照一定规则收集、整理、标注并打包好的一组数据的集合。高质量数据集可达到随手即用的程度,直接被用于开发和训练AI(机器学习)模型,能有效提升模型性能^[12]。训练不同模型所需要的数据集也有所不同。

机器学习是建立数据驱动材料专用模型的基本方法,相应使用的数据集聚焦于特定主题。在训练通用性更高的AI模型时,数据集的内容需要更综合和多样化,以支持多模态数据的融合和复杂系统的建模。例如智能医疗中,可能需要结合医学影像、基因表达和临床病历数据集;在开发智能化系统时,需要动态更新的数据集,以适应不断变化的环境和任务需求。例如,智能实验室需要实时更新的数据以不断优化实验计划^[17-18]。此时,使用的数据集中除既有数据外,也包括实时动态更新的数据。构建此类高质量材料数据集^[12]所需的素材,由数据源中的历史数据(数据库)和实时数据(实验、计算装置新产生的数据)共同组成^[19]。

语言模型是AI模型中重要的一个类别。大语言模型(large language model, LLM)是指基于深度学习、拥有大量参数(通常为10亿级以上)、在海量文本数据上进行训练的自然语言处理(natural

language processing, NLP)模型^[20],它的目标是让机器理解语言、生成语言、参与任务型语言交互。近年来,针对特定应用场景优化的专业化大型语言模型(Specialized LLMs)快速涌现,反映了从通用AI向垂直领域深度定制解决方案的重大转变,小型化与专业化语言模型被认为将引领AI未来的发展^[21-22]。在训练专业化语言模型时除多模态数据集外,还需要使用领域特定语料库。语料库(Corpus)是将在语言的实际使用中真实出现过的语言材料经过系统采集、清洗、标注、归档、持续维护的“机器可读的电子文本图书馆”^[23]——有规范、有元数据、有版本管理,可检索、可复用、可引用,没有它NLP和AI技术就无法施展。

数据集、语料库都是AI-ready数据在应用层面的表现形式。从内容的角度看,任何数据集(或语料库)都只是从材料全部信息(或数据全集)中选取的一个子集。就好比一个人的平面照片仅反映了此人从特定角度看去的样貌,完整的形象需要从全方位对其进行观察、拍摄。数据集建立的另一个维度是数据覆盖范围,即哪些条件下产生的数据应该被组织进来。数据集(或语料库)是构建者对于特定目标的科学思维与认识的反映,因此即使针对同一主题(如锂离子电池)也不是唯一的。满足一位研究者需求的数据集,未必满足另一位研究者;当前有效的数据集,未来可能不一定适用。

传统数据集的局限性,还包括习惯性地缺乏实验设置的细节信息、没有得到发表的数据和负数

表1 不同场景下材料数据的用途、目的、作用、形式与特点

用途	目的	作用	形式	参数量	参数特点
材料计算	特定参数	事实依据	单条数据	单一	样品-性质,单一模态
机器学习建模	专用模型	信息元素	数据集	多维度	聚焦于主题多模态
人工智能建模	通用模型	信息元素	数据集	多维度	综合和多样化,多模态
大语言模型	语言模型	信息元素	语料库	多维度	语言表达的数据

据(即不尽人意的结果数据)等。大量富有意义的实验细节常被熟练的数据采集者看成是细枝末节而熟视无睹,造成在使用时因细节缺失而留下大量不确定性。很多数据可能由于平淡无奇没有被写入文章发表,但填充了参数空间的空缺;还由于多年来研究是以“结果为王”,造成结果不理想的数据被认为是“坏数据”。这些有益却时常缺席数据集的数据被统称为“暗数据”^[3]。它们对于机器学习的独特价值也尚未被充分认识。实际上,“暗数据”所提供的是更高的数据维度、更宽阔的参数空间覆盖与更均衡的参数分布。解锁暗数据的潜力,有望在开发精确且稳健的科学 AI 模型中发挥至关重要的作用^[3]。

因此,要全面满足 AI 对材料数据的需求,有必要建立材料数据全集,即所有材料的所有信息。对每种材料来说,包含完整描述与其相关的制备全部过程及参数、对其施行的全面表征的原始数据(未经处理的),以及对数据分析处理得到的材料特性参数结果,或称之为全信息数据,见第 2 章节。可以将材料数据全集想象成一个汇聚各种材料全信息数据的池子(这里

用“池”意在与数据“湖”和数据“仓库”形成区别),其中收纳了构建所有(不是一种)数据集必要的信息元素;与此同时,覆盖了宽阔的参数空间的材料样本被无差别地收录。这当然是一种理想化的情况,在实际中数据库/资源节点也是根据一定主题设置的^[13-15],不可能应有尽有。能够操作的是在可能条件下,本着“能采尽采”原则,通过长期积累,逐渐接近“数据全集”。全信息数据池是数据集之源,提供比数据集更全面、更详尽、更深层次的材料数据基础资源。

如上所述,材料数据基础设施的资源可分为 2 个层面:一是实用资源,包括数据集和语料库,直接面向当前 AI 建模需求;二是基础资源,提供构建数据集和语料库所需的数据源,体现为全信息数据池,即能够提供“全信息数据”的原始数据集合。二者的关系可类比为一棵树(图 1),数据源层面的资源就好比是树根,称之为根资源;数据集层面就好比是枝叶,称之为叶资源。根深才能叶茂,固本才能荣枝。枝叶靠树根的养分得以繁茂,缺乏了根资源则叶资源就无法生长出新的枝叶,难以支撑长

期发展的需要。

当前,材料研发智能体被认为是 AI4M 的重要发展趋势,其研发对于高精度的语料库极其渴望。语料库是以语言表达的数据集,形式上表现为文本集合。专业语料库一般是通过收集大量相关专业领域文献资料,提取相关的文本、段落和句子构建而成,资源有限且质量参差不齐。现实中有大量实验结果在文献中不能得到精准而细致的描述,同样的实验也会因作者的关注点各异或写作习惯不同,造成语料一致性的偏差。这个问题可以通过建设标准化“语料加工生产线”得到解决。在全信息数据池中,存储了大量的材料制备过程的详细、完整的标准化结构化数据,以及对这些材料开展的各种表征数据。这些数据采集是基于标准化流程建立的数据模板(第 3 章节),其中隐含了具有严格定义的过程内涵,因此可以方便地通过计算机代码以标准形式批量化地转换为高精度语料。这为快速产出大批量高质量语料提供了新途径,对于材料大模型和智能体的开发具有很高价值。Ashok 等^[24]的研究指出,少量人工标注数据点(如 200 个)为模型带来的性能提升效果,相当于增加 2~3 个数量级的额外合成数据点。这或许预示着在高精度语料条件下,训练语言模型可以在少得多的数据量下进行,让我们有了用“精度换数量”的期许。

2 模块化可组装材料数据模型

在建设数据源层面的数据资

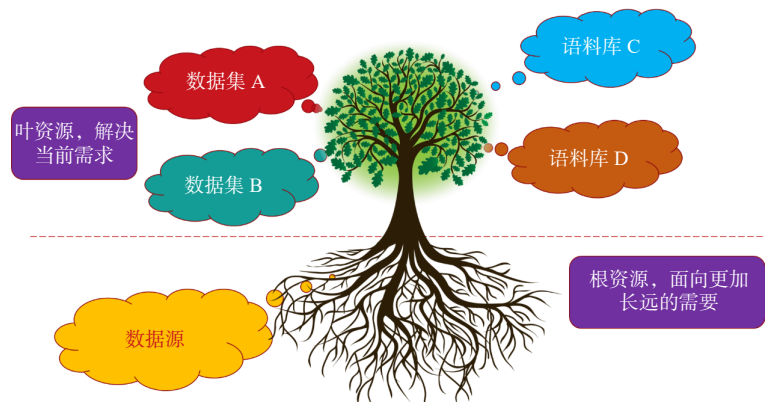


图 1 数据资源树

源时,全信息数据池的数据标准化非常重要,通过对同类原始数据形式、格式、内容设定统一的标准规范,不仅可以确保数据的共享流通,更重要的是通过结构化建模和语义标注,实现数据资源的机器可读性与自动化处理能力,为智能算法的自主学习和深度分析提供高质量数据支撑。类似于数据仓库,标准化需在数据入池前就完成(Schema-on-Write),缺少了该标准化流程,全信息数据池就会变成无法查询、无法融合、无法信任的“数据迷宫”。但数据处理要等到使用时才根据需要进行,这点又类似于数据湖(湖中的数据未经标准化,待使用时才进行必要的处理)。数据标准化的另一项重要意义是对数据的内容进行规范,避免不同机构、不同个人、不同实验初衷所开展的同一实验产生出各具个性的数据,这对于全信息数据的采收尤其重要。

图2为一个假想材料的全信息数据内容示意图。左侧列中包含了材料合成制备的全过程,通常由多步可独立进行的制备工艺所组成(单一步骤是特例),每一步骤均涉及复杂的动态过程,其过程的参数对于材料的最终性能有着决

定性的影响(如一块合金样品可能经历熔制、冷轧、退火等过程)。因此,材料全信息数据需完整涵盖各制备步骤的相关数据。与此同时,材料样品需通过多种表征/测量手段(如X射线衍射分析、扫描电子显微镜、背向散射电子衍射技术、拉伸试验等)才能完整地获得研发所需的特性/表现数据,且在研发不同材料类型时所用到的表征/测量手段也存在差异。这里还需要强调一点,全信息数据池中收集的表征数据是未经处理的原始数据,多数情况下,材料的特性是需要对表征原始数据进行处理才能得到的(个别情况下,表征原始数据直接反映了材料的特性)。传统材料数据库一般仅收集由原始数据处理而得到的分析结果(如材料性能参数等),构成简单对应关系,这虽然方便了使用,但限制了从同一数据中继续深入挖掘出更多信息的可能性。图2最右侧列是基于对数据分析处理得到的材料特性参数,称之为衍生数据,这是通常建立数据集所需的。一种特性可能从一种表征数据推出,也可能从多种表征结果得出。

材料数据具有不同于一般数据的某些特征。首先,每一样品的

数据都涉及多个维度,如成分维度(主元素、掺杂、杂质含量、同位素等)、结构维度(相、晶系、缺陷密度、织构指数等)、性能维度(力学、热学、电化学、辐照、海洋环境条件等),难以用一套标准模板适用多类材料。其次,材料数据总是与条件耦合在一起的,温度、湿度、应变速率、气氛、取样方向、热处理路线等均与最终性能密切相关(有制备也有表征、服役方面的因素),例如同一块钢在25℃与-40℃下的冲击功可以相差1倍^[25]。因此,采集详尽的制备过程元数据与表征条件元数据都是必不可少的。特别是在研发中表征种类往往随需要不断增加,无法预先确定。数据的内容(参数量)发生动态变化,也就难以制定统一规定。综上所述,这种数据参量的多样性与多变性造成对全信息材料数据进行标准化非常困难。

观察图2可以得出,复杂的材料的制备与表征过程可拆解为多个独立完成步骤的组合。基于此,提出一种模块化可组装数据模型,其要义主要有3点:(1)以可独立完成的制备、表征、衍生过程作为数据采集、存储的模块单位,完整的材料全信息数据由相应数据模块组合而成;(2)每个模块代表一个具有共性的、可重复进行的过程,其内容有着清晰定义的内涵,意味着过程可以用标准形式进行语言描述,具备标准化语料生成的条件;(3)为每个数据模块分别建立标准,通过全面的局部(模块)标准化,实现材料数据整体的标准化。这种化整为零的方式大大降低了复杂数据的标准化难度。材

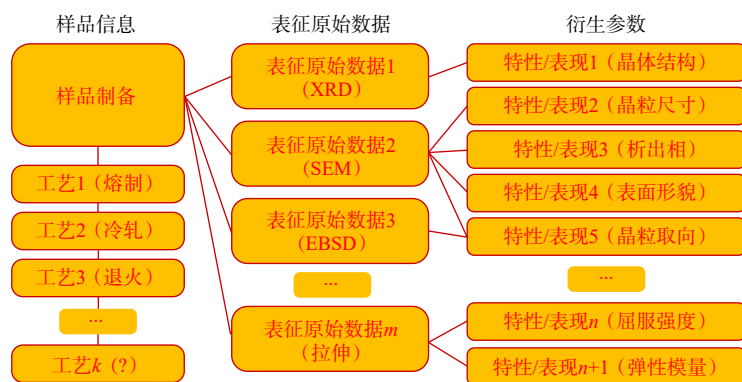


图2 全信息数据内容示意(由此发展出模块化可组装材料数据模型)

料原始数据标准化的关键在于对材料制备和表征过程的参数化表达的标准化,而非所有的数据条目都遵从整齐划一的格式。因此无需担心不同样品的全信息数据具有不同长度。

以上全信息数据的示例是通过基于实验方法的制备与表征进行说明的。在实际中,样品既可为实验产生的实物,也可为经计算产生的虚拟物。表征原始数据既可来自于实际表征或测量,也可通过模拟计算产生。其处理方法和逻辑与实验完全一样。

为进一步说明,将模块化可组装数据模型与 LEGO(乐高)玩具进行类比^[26]。从店里买到的 LEGO 玩具是一大盒标准化的砌块,用它们可以搭建出各式各样的物件(图 3^[26])。因为每一个砌块都是标准化的,搭出来的物件也就是标准化砌块的组合,自然可以用标准化的语言来描述。一个材料的全信息数据可类比为一个搭好的乐高玩具,每个砌块代表一个数据模块,表示一条可独立完成动作的数据。那么,每个数据的标准化模板就隐含了对动作过程及其参数的官方定义,由这些模块的集合所表达的物件自然也都是使用标准的词汇进行描述,不会出现“生词”与非标的表述。

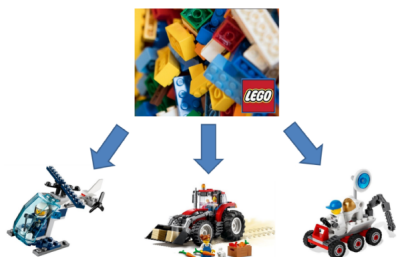


图3 LEGO玩具(用标准化的砌块搭出丰富多彩的物件)

当前 2 种常用的数据存储架构是数据湖与数据仓库。数据湖以数据产生时的原始格式存储任意类型数据,大幅降低采集存储成本,为 AI/机器学习提供灵活数据基础。但其缺点除使用不便外,在数据来源涉及多地点、多仪器的情况下,存在数据内容与质量不确定等问题。数据仓库则采用预定义方式,存储经过清洗和结构化的数据,如数据集,但难以应对爆发增长的非结构化数据^[27]。在为一个机构设立的自用数据设施中,经常通过湖仓一体融合两者优势,既保留数据湖的原始数据存储能力与 AI 应用潜力,又具备数据仓库的结构化报告与数据治理功能。然而,对于具有公益与公共属性的材料数据基础设施,需要实现国内各机构间,乃至国际间的数据集与原始数据流通,原始数据必须标准化,使得在不同地点不同时间获得的同一类原始数据能够对齐,进行批量处理,同时又能保持最大限度的灵活性,这使得数据湖与数据仓库都难以胜任。因此,最为理想的情况是采用“池仓一体”在当下与长远之间找到平衡。如果说数据湖是存放原材料的堆场,数据仓库就是摆放成品的货架,而全信息数据池则更像一个存放标准化零部件的仓库。当前最为急需的是货架上的成品,而未来将越来越多地使用标准化的部件组装出新的成品。

3 模块化可组装数据标准化原则

中国在材料基因工程方法兴

起的初期,就认识到数据标准的重要性。2017 年,在中国材料试验标准委员会(CSTM)成立之时,王海舟院士前瞻性地提出组建 CSTM 材料基因领域标准化委员会,这是国际上第一个材料基因工程领域的标准委员会,率先开展材料基因工程领域标准与标准化的重要探索与示范。2017 年 11 月,在第一届材料基因工程高层论坛期间,CSTM 材料基因领域标准化委员会(CSTM FC-97)正式成立,下设通则、计算、制备、表征、数据、应用 6 个技术委员会。考虑到材料基因工程以“数据+AI”为核心的特点,确定将围绕数据展开相关标准制定。FC-97 提出的材料数据标准化总体建设方针是:(1) 设计符合材料科学与 AI 特点的材料数据结构(如模块化可组装数据结构);(2) 基于这个数据结构,制定出与此相适应的材料数据通用规则;(3) 通过顶层设计,建立面向未来的、适合以上材料数据系统的数据标准体系;(4) 基于数据标准体系,有序地、批量化地制定一系列具体材料数据标准与规则。

在模块化可组装数据结构的基础上,FC-97 从建立数据通用规则入手,在充分满足科学数据可发现、可访问、可互操作、可重用(FAIR)原则的前提下,首先确立数据条目中所包含内容的原则。2019 年 8 月,CSTM 发布了由国内 30 余家材料研究主体单位共同制定的《材料基因工程数据通则》T/CSTM 00120(图 4,简称《通则》)^[28],《通则》针对全信息数据的内容,将数据分为样品信息(制备数据)、原始数据(未经处理的表征数据)与衍生数

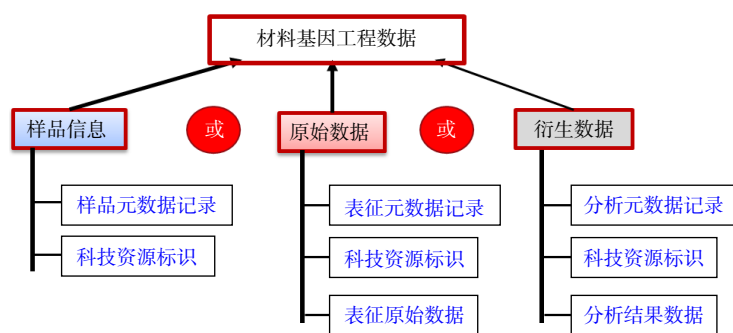


图4 材料基因工程数据通则内容图示

据(经分析处理得到的数据)3类。其中,样品既可为实验产生的实物,也可为经计算产生的虚拟物。原始数据既可来自于表征或测量,也可通过模拟计算产生。

《通则》是面向全信息数据而设计的,重点解决3个问题:(1)适配模块化可组装数据结构,规定每条数据以可独立完成的操作(样品制备/表征/计算/数据处理)为模块单位,仅包括与该操作相关的内容;(2)元数据元素中需包含足够信息描述样品产生、表征和数据处理过程,遵循“能收尽收”原则;(3)表征原始数据(未经分析处理的数据)得到采集。原始数据中包含大量的信息,其为不同的目的多次利用是数据可再利用性的重要体现。目前原始数据大多分散在产生者手中,未被收录,极大地限制了数据的再利用。该分类方式从制度上确保原始数据被记录下来,从而保证了被深层次利用的可能。此外,为满足FAIR原则^[29],对每条数据均分别赋予独立且永久的资源标识,例如,依据国家标准GB/T 32843等规则^[30],也可依据任何独立赋予的唯一且永久的标识体系如数字对象唯一标识符(digital object identifier, DOI)^[31]。将样品单独列为一类数据是一种

全新的做法,这意味着样品本身成为符合FAIR原则的公共社会资源,便于建立与样品信息数据平行的有实体样品库,有利于实体样品的共享、多用和重复使用。

《通则》为材料全信息数据的标准化建立了基点。在《通则》的原则指导下制定的《材料基因工程数据元数据标准化原则与方法》T/CSTM 00837^[32],为材料数据模块化数据标准的制定提供了实际操作方法。元数据的完整性是数据再利用的基础条件,但当前材料数据收录的元数据普遍存在缺失。因此,该文件明确规定,在具体标准制定中必须本着“能收尽收”原则,收集足够的元数据,以保证数据可被他人充分利用。

4 讨论与总结

在AI时代,材料科学已从传统的“经验试错”跃升为“数据驱动”的新范式。决定AI4M能否持续进化的关键,不再是算法,而是“数据土壤”,即一套可被机器随手取用的、标准化、可共享、带完整上下文的全信息数据体系。材料数据资源可分为2个层面:一是实用资源,包括数据集和语料库,直接面向当前AI建模需求;二是

基础资源,提供构建数据集和语料库所需的数据源,体现为全信息数据池。当前各国都把高质量数据集作为建设材料数据资源基础设施的核心内容,中国正在建设中的新材料大数据中心拟围绕主平台建立一系列资源节点^[10],聚焦特定材料领域解决高质量数据集不足的问题。笔者认为在着力发展高质量数据集的同时,数据源层面的内容应同样受到重视。面向未来的数据基础设施必须协同构建“根资源”与“叶资源”。其中根资源(全信息数据池)是保障数据可溯源、可再利用、可长出“新枝叶”的基础。对于具有公益与公共属性的数据基础设施,必须对根资源实行数据标准化。本文提出了一种模块化可组装的材料数据模型及与此相适应的数据标准化原则,这使得数据标准化可方便地应用于灵活多变的样品制备、表征和数据衍生技术组合,大大降低了复杂数据的标准化难度。

尽管国内外材料大数据领域在技术、标准、资源汇集方面已取得一些进展,但均存在较明显的难点与痛点。(1)体系化的材料数据标准普遍尚未健全,数据碎片化与“孤岛化”现象仍然突出,导致跨平台数据融通困难;(2)数据共享生态尚未形成,科研单位和企业对数据积累与共享的积极性不够,公共数据开放共享范围有限,导致数据流转不畅、开发利用不深入;(3)兼具材料科学、数据技术与标准技术的复合型人才缺乏,从事标准研究与制定的人员数量严重不足,难以支撑国家全面建立材料大数据标准的需要,进一步制约了技

术突破与 AI-ready 数据生态的构建; (4) 限于资源配置, 标准制定工作安排较为分散, 缺少集中的应用平台, 难以开展系统性的标准化示范与效果评估。值得欣慰的是, 随着新材料大数据中心建设的全面开展, 通过顶层设计, 各项工作正在有计划, 有针对性地进行部署, 上述问题有望系统性地得到逐步解决, 前景可期。

新材料大数据数据资源基础设施的建设是一项长期性的事业, 关系到中国新材料领域的长远发展。作为国家材料数据的核心汇聚平台, 在建设须做好前瞻性的规划, 既立足当下, 做好数据集, 解决眼前的急需, 更要放眼未来, 做好数据源, 为未来的发展留出足够空间, 满足 AI 日益增长的数据需求。因此, 具有公益与公共属性的数据基础设施应具有“池仓一体”的构架, 兼顾数据集与全信息数据池, 平衡当下与长远。当全信息数据池成为重大材料数据基础设施的核心部分, 且模块化标准覆盖所有材料制备与表征环节时, 中国将不仅拥有全球最大的材料数据库, 更将拥有全球最完整、最鲜活、最可持续生长的“材料数据生态”, 为中国抢占未来产业制高点提供不竭的“数据赋能”。

致谢: 在本文撰写与修订过程中, 同多位专家学者进行了深入而富有启发性的探讨, 包括北京科技大学/苏州国家实验室宿彦京教授、北京科技大学张雷教授、上海大学施思齐教授、中国科学院上海硅酸盐研究所刘建军研究员、苏州国家实验室张坦副研究员, 在此深表感谢。

参考文献(References)

- [1] 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.
- [2] The White House. Launching the genesis mission[EB/OL]. (2025–11–24) [2026–02–24]. <https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/>.
- [3] GOV of UK. AI for Science Strategy [EB/OL]. (2025–11–20) [2026–02–24]. <https://www.gov.uk/government/publications/ai-for-science-strategy/ai-for-science-strategy>.
- [4] 中华人民共和国中央人民政府. 国务院关于印发新一代人工智能发展规划的通知[EB/OL]. (2017–07–20) [2026–02–24]. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.
- [5] 王毅, 李高楠, 刘哲, 等. 材料基因工程与智能科学: AI+时代无尽前沿[J]. 科技导报, 2025, 43(12): 93–109.
- [6] 宿彦京, 付华栋, 白洋, 等. 中国材料基因工程研究进展[J]. 金属学报, 2020, 56(10): 1313–1323.
- [7] 中华人民共和国中央人民政府. 国务院办公厅关于印发科学数据管理办法的通知[EB/OL](2018–03–17) http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.
- [8] 郭华东. 科学大数据: 国家大数据战略的基石[J]. 中国科学院院刊, 2018, 33(8): 768–773.
- [9] 高孟绪, 王瑞丹, 王超, 等. 关于国家科学数据中心建设与发展的思考[J]. 农业大数据学报, 2019, 1(3): 21–27.
- [10] Wang H, Xiang X D, Zhang L T. On the data-driven materials innovation infrastructure[J]. *Engineering*, 2020, 6(6): 609–611.
- [11] 工业和信息化部, 财政部, 国家数据局. 新材料大数据中心总体建设方案[EB/OL]. (2024–10–16) [2026–02–24]. https://www.gov.cn/zhengce/zhengceku/202410/content_6983916.htm.
- [12] 全国数据标准化技术委员会. 高质量数据集建设指南[EB/OL]. (2025–08–29) [2026–02–24]. <https://www.tc609.org.cn/tc609/tzgg/202509/6d0f135392ca47dfa28dad22f7bb5f6b.shtml>.
- [13] 王冠杰, 李开旗, 彭力宇, 等. 高通量自动流程集成计算与数据管理智能平台及其在合金设计中的应用[J]. 金属学报, 2022, 58(1): 75–88.
- [14] 孙志梅, 王冠杰, 张烜广, 等. 材料基因工程加速新材料设计与研发[J]. 北京航空航天大学学报, 2022, 48(9): 1575–1588.
- [15] Wang G J, Peng L Y, Li K Q, et al. ALKEMIE: An intelligent computational platform for accelerating materials discovery and design[J]. *Computational Materials Science*, 2021, 186: 110064.
- [16] 李振华, 王同益. 数据治理[M]. 北京: 中共中央党校出版社, 2021.
- [17] Zhu Q, Zhang F, Huang Y, et al. An all-round AI-chemist with a scientific mind[J]. *National Science Review*, 2022, 9(10): nwac190.
- [18] Dai T W, Vijayakrishnan S, Szczyński F T, et al. Autonomous mobile robots for exploratory synthetic chemistry[J]. *Nature*, 2024, 635 (8040): 890–897.
- [19] 张澜庭, 刘明洋, 袁宏建, 等. AI驱动的新材料智能研发与数据标准化[J]. 科学通报, 2025, 70(24): 4066–4080.
- [20] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[J]. *Advances in neural information processing Systems*, 2020(33): 1877–1901.
- [21] Touvron H, Martin L, Stone K, et al. LLaMA: Open and efficient foundation language models[J]. arXiv preprint arXiv: 2302.13971(2023).
- [22] Yang C, Zhao R, Liu Y, et al. Survey of Specialized Large Language Model[J]. arXiv preprint arXiv:

- 2508.19667 (2025).
- [23] O'effe A, McCarthy M J. The routledge handbook of corpus linguistics[M]. 2nd ed. London: Routledge, 2022.
- [24] Ashok D, May J. A little human data goes a long way [EB/OL]. [2025-12-10]. <https://arxiv.org/abs/2410.13098>.
- [25] 白泉, 盛国华, 金生吉. 钢结构设计原理[M]. 北京: 清华大学出版社, 2024.
- [26] 乐高(LEGO)公司. [2025-12-10]. <https://www.lego.com/zh-cn>.
- [27] 思爱普(SAP)公司. (2025-11-12) [2025-12-10]. <https://www.sap.cn/resources/data-lake-vs-data-warehouse>.
- [28] 中关村材料试验技术联盟. T/CSTM 00120—2019, 材料基因工程数据通则[S]. 北京: 中关村材料试验技术联盟, 2019.
- [29] Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR Guiding Principles for scientific data management and stewardship[J]. *Scientific Data*, 2016, 3: 160018.
- [30] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB/T 32843—2016, 科技资源标识[S]. 北京: 中国标准出版社, 2016.
- [31] Liu J. Digital object identifier (DOI) and DOI services: An overview[J]. *Libri*, 2021, 71(4): 349-360.
- [32] 中关村材料试验技术联盟, T/CSTM 00837—2022, 材料基因工程数据元数据标准化原则与方法[S]. 北京: 冶金工业出版社, 2022.

Artificial intelligence-oriented materials data infrastructure and data standardization

WANG Hong^{1,2,3,4}

1. School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
2. Suzhou National Laboratory, Suzhou 215123, China
3. Materials Genome Initiative Center, Shanghai Jiao Tong University, Shanghai 200240, China
4. Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai 201203, China

Abstract Artificial intelligence is driving materials research toward a new data-driven paradigm of AI for Materials (AI4M), with data being a key foundation for the success. Both the Genesis Mission program of U.S. and the AI for Science Strategy of U.K. identify the collaborative operation of high-quality research datasets and supercomputing as national strategic infrastructure. China's Overall Construction Plan for the New Materials Big Data Center also clearly aims to integrate national data resources to accelerate materials research using AI. Currently, the lack of high-quality, AI compatible "ready-to-use" data is considered as the biggest bottleneck in AI4M. It should be recognized that datasets are only a component of a larger whole of materials information since any dataset merely reflects the scientific understanding of its creators. Limitations and bias are inevitable. The datasets that are effective today may not necessarily be so in the future. This paper systematically analyzes new requirements that AI imposes on the content and form of materials data, and proposes a future-oriented materials data infrastructure. It should be structured like a "tree" with a comprehensive information data pool as the root and high-quality datasets/corpora as the leaves, which allows for sustainable evolution. By standardizing root resources in advance, raw data can be reused and repurposed to meet the evolving needs of specialized models and general large language models. A modular and assemblable data model is designed to address the standardization challenges caused by the diversity and variability of material data. Building a national-level materials big data infrastructure is a long-term endeavor, crucial to the future development of China's new materials industry. Forward-looking planning is required—not only to meet the current needs of creating quality datasets, but more importantly, to aim on potential future demand on data sources, leaving enough room for development in the years to come. Therefore, material data infrastructures with public welfare and service attributes should have an 'integrated pool and warehouse' structure, catering to both datasets and comprehensive materials information pools, to balance the present with the future.

Keywords data infrastructure; data sets; data sources; data standardization; comprehensive materials information data pool ●



(责任编辑 魏来)