材料基因工程数据生态系统

尹海清[1,2]，王永伟[2]，张晓彤[1]，姜 雪[1]，黄海友[1]，张 雷[1]，
岩田修一[3]，宋晓艳[4]，张洪梅[5]，姚磊江[6]，崔丽娜[7]，
周 勇[8]，陈 宁[1]，曲选辉[1]

（1. 北京科技大学 北京材料基因工程高精尖创新中心，北京 100083）
（2. 北京科技大学 钢铁共性技术协同创新中心，北京 100083）
（3. 日本东京大学理学部，东京 113-8654）
（4. 北京工业大学材料科学与工程学院，北京 100124）
（5. 北京理工大学材料学院，北京 100081）
（6. 西北工业大学 无人机特种技术国防科技重点实验室，陕西 西安 710129）
（7. 中国科学院金属研究所 沈阳材料科学重点实验室，辽宁 沈阳 110016）
（8. 中国科学院化学研究所 中国科学院工程塑料重点实验室，北京 100190）

摘 要：在全球数字化飞速发展的时代，互联网、物联网和计算机技术使数据的获取、存储和分析变得非常高效便捷。数据科学作为科学发现的第四范式，进一步加快了数据在科学研究中的应用。政府、企业、科研院所甚至个人都积累了大量数据，越来越多的人充分认识到数据对社会经济和科技的推动作用，因此对数据生态系统的需求应运而生，需要从全生命周期的角度来构建从数据的生产到输出、应用的全过程的生态系统。对材料数据生态系统的七要素：材料数据分类、材料数据标准、数据采集、数据存储、材料数据知识产权保护、数据质量管控以及材料数据库的重要性和最新进展进行论述，描述了材料基因工程数据生态系统的建设情况。

关键词：数据分类；数据标准；数据知识产权；数据生态系统；材料基因工程；全生命周期
中图分类号：TP311.13 **文献标识码**：A **文章编号**：1674-3962(2023)02-0135-09
引用格式：尹海清，王永伟，张晓彤，等. 材料基因工程数据生态系统[J]. 中国材料进展，2023，42(2)：135-143.
YIN H Q，WANG Y W，ZHANG X T，*et al*. Material Data Ecosystem for Materials Genome Engineering[J]. Materials China，2023，42(2)：135-143.

Material Data Ecosystem for Materials Genome Engineering

YIN Haiqing[1,2], WANG Yongwei[2], ZHANG Xiaotong[1], JIANG Xue[1], HUANG Haiyou[1], ZHANG Lei[1], IWATA Shuichi[3], SONG Xiaoyan[4], ZHANG Hongmei[5], YAO Leijiang[6], CUI Lina[7], ZHOU Yong[8], CHEN Ning[1], QU Xuanhui[1]

(1. Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing 100083, China)
(2. Collaborative Innovation Center of Steel Technology, University of Science and Technology Beijing, Beijing 100083, China)
(3. School of Science, The University of Tokyo, Tokyo 113-8654, Japan)
(4. School of Materials Science and Engineering, Beijing University of Technology, Beijing 100124, China)
(5. School of Materials Science and Engineering, Beijing Institute of Technology, Beijing 100081, China)
(6. National Key Laboratory of Science and Technology on UAV, Northwestern Polytechnical University, Xi'an 710129, China)
(7. Shenyang Key Laboratory of Materials Science, Institute of Metal

收稿日期：2021-12-01 **修回日期**：2022-04-30
基金项目：国家重点研发计划项目(2020YFB0704504，2016YFB0700503)
第一作者：尹海清，女，1971 年生，教授，博士生导师，Email：hqyin@ustb.edu.cn
DOI：10.7502/j.issn.1674-3962.202112001

Research, Chinese Academy of Sciences, Shenyang 110016, China)
(8. Key Laboratory of Engineering Plastics, Institute of Chemistry, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In the era of rapid global digitalization, the internet, the Internet of Things (IoT) and computer technologies have made it highly efficient for data acquisition, storage and analysis. Data science, as the fourth paradigm of scientific discovery, has further accelerated the application of data in scientific researches. Governments, enterprises, research institutes and even individuals have accumulated a large amount of data. The role of data played in promoting the development of society, economy and technology is fully understood, therefore, there rises the demand for the data ecosystem, covering the whole production line from data production to data application. In this case, the necessity and progress of the essential features of data ecosystem are elucidated, including data classification, data standards, data collection and storage, intellectual property protection, data quality control and database of material data. The initial goal is to identify the significance of material data ecosystem. And the framework of data ecosystem of materials genome engineering is briefly demonstrated.

Key words: data classification; data standards; data intellectual property; data ecosystem; materials genome engineering; life cycle

1 前 言

生态系统的概念是由英国生态学家 Sir Arthur George Tansley 博士在 1935 年提出，即由有机复合体和形成环境的整个物理要素共同组成的一个物理系统，这种系统具有多种大小和种类，覆盖从宇宙整体到原子的范围。关于数据生态系统最早的两篇文献于 2011 年发表，Tim Davies[1] 提出数据生态系统由数据基础设施和标准组成，以鼓励数据开放共享及相关新技术开发；Ding 等[2] 构建了支持政府数据开放的平台，平台与参与者一起构成了一个可以交互的生态系统。对数据生态系统的定义众说纷纭，目前尚未形成一个统一且明确的描述。Oliveira 等[3] 将参与者及其角色、关系及数据资源 4 个要素相结合，定义数据生态系统由一组松散的交互参与者组成，他们直接或间接地使用或产生数据及其他软件、服务和基础设施等相关资源。数据生态系统聚焦数据，由社会与技术相关联组成一个复杂网络，生产、检索、存储、发布、使用及重用数据并促进创新和创造新价值[4]。数字化的进程使不同领域与行业开始寻求数字化转变，例如能源、环境等领域，其中数据交互、基于数据的预测与决策等关键问题都需要数据生态系统的支撑。Anwar 等[5] 将能量数据生态系统的组成分为 6 个方面，即参与者、能力(capability)模型、系统支撑服务、知识管理模型、基础设施以及数据。在数据生态系统的发展中，安全、价值链等也成为了核心问题[5, 6]。

材料的发展水平直接决定了一个国家的工业化基础。科研人员对材料数据的研究和应用已有至少 20 余年的历史，集成计算材料工程(integrated computational materials engineering, ICME)[7]、材料基因组计划(Materials Genome Initiative, 我国称之为材料基因工程)[8] 在 2008 年及 2011 年由美国相继提出，将材料研究进一步引向了数据驱动的创新研究方向，吸引了国内外大量研究人员开展相应工作。美国宾州州立大学刘梓葵教授[9] 针对热力学计算及相关研究工作产生的数据，提出基于可扩展的、自我优化的相平衡基础设施 ESPEI 的数据可持续生态系统的“数据海”的概念，系统包括数据存储库、互联、私有数据、数据处理、数据收集及数据循环使用，数据海的描述为材料数据生态系统的构建提供了技术指导。

我国近几年，尤其是“十三五”期间，一批重点研发计划项目等国家及地方项目的投入也使得数据及其分析应用成为材料研究热点，材料基因工程项目支持的材料数据库的建设纷纷落地。为了使材料数据发挥更大的效力，避免大数据时代的数据孤岛的出现，本文作者基于材料数据共享网与材料基因工程数据库建设，以及数据挖掘融合及其在材料基因工程中的成果[10-12]，探索材料数据生态系统的构建思想，为数据驱动的材料研发与材料智能制造提供数据理论支撑。

材料数据生态系统架构的思想不仅包括材料数据本质的、全生命周期的内涵式发展，同时外延至材料数据活动的参与者以及与数据资源的复杂关系，两个维度的发展与交互构成了相对完整的材料数据生态系统。图 1 为材料数据生态系统中全生命周期的特征描述，包括数据分类体系、数据库建设、数据标准、数据采集、数据存储、数据共享及数据质量评估 7 个特征。

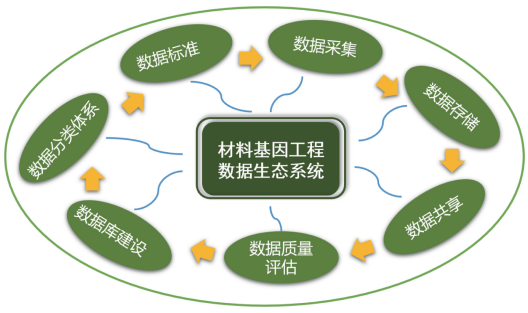

图 1 材料基因工程数据生态系统中的数据全生命周期特征描述

Fig. 1 The life cycle data features demonstration of the data ecosystem of Materials Genome Engineering (MGE)

2　材料数据分类体系的建立

当数据科学作为科学研究的第四范式与材料研究领域相结合时，数据被赋予了材料的属性，材料数据科学也成为了材料研究的一个分支，而数据分类成为材料数据科学的基础之一。由李依依院士任组长、院士及顶级专家共 11 位组成的专家组依据师昌绪院士主编的《材料大辞典》中的材料分类体系，构建了材料数据分类体系，该体系是目前国际上唯一的材料数据分类体系[13]，如图 2 所示。该分类体系包括九类一级材料数据类别，如图 2a 所示，包括材料基础数据、金属材料数据、无机非金属材料数据、高分子材料数据、复合材料数据、信息材料数据、能源材料数据、生物医用材料数据、天然材料及其制品数据等，其中材料基础数据包括纯元素、二元及多元材料的原子尺度的特征及微观结构的热动力学数据等。一级数据类别下包含二级、三级等多级材料数据分类，全面覆盖了材料各个领域，图 2b 为金属材料数据一级目录下的二、三级目录。

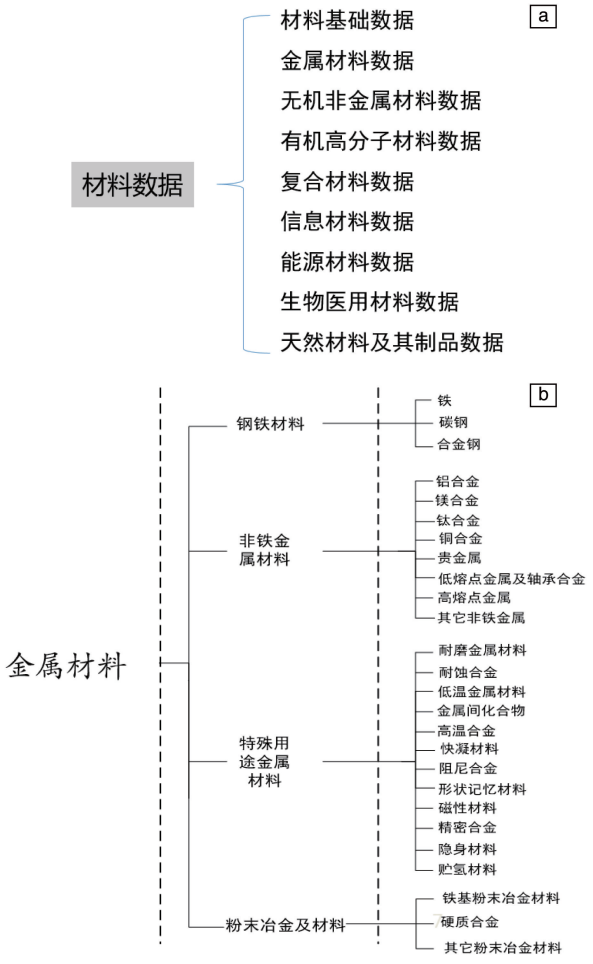

图 2　材料数据分类体系的第一级目录（a）及金属材料数据的一级、二级分类目录（b），其中材料名称后省略“数据”二字

Fig. 2　The first level of the material data classification system (a) and the primary, and secondary catalogs of metal material data classification (b), the word “data” is omitted here

图 2 的材料数据分类体系在提出后的十余年时间里得到了国内外材料数据研究者的认可，在科技部科技基础条件平台项目支持建设的国家级数据库材料科学数据共享网[14]中得到应用，并在“十三五”国家重点研发计划项目支持的材料基因工程数据库的建设中进一步得到应用，形成材料数据库的科学架构。

材料数据的获取途径包括实验与计算。计算数据和实验数据在元数据、精度及应用等方面存在着较大的差异。数据挖掘与机器学习获得的数据，依据获取过程的本质，将它们归为计算数据一类。计算数据按描述的尺度可以进一步细分为第一性原理、分子动力学、热力学、动力学、宏观有限元计算等多类计算数据。但无论哪一尺度的计算，其计算对象都属于某一类材料，因此，按照数据获取途径的数据分类服从于材料数据类型分类。

3　材料科学数据标准系统

2021 年 10 月，国务院印发《国家标准化发展纲要》，进一步强调了标准化在推动高质量发展中的基础性和引领性作用，并提出了经济全域、国内国际相促进以及质量效益型的新发展方向。国家标准《标准化工作指南 第 1 部分：标准化和相关活动的通用词汇》中指出，标准化的主要作用是针对一个或多个特定目的，使产品、过程或服务适合其用途，防止贸易壁垒，并促进技术合作。标准已经成为目前各行业全球竞争的热点，更是成为国际竞争中挤掉竞争对手的工具，拥有了制定产品标准的话语权往往就意味着对产品市场的控制权。在全球数字化飞速发展的时代，标准的作用远不仅限于此。

随着材料基因工程相关项目的实施，众多材料科研人员投入到材料数据库建设与数据驱动的材料创新研究中，材料数据标准直接关乎到数据的质量和数据挖掘应用的基础，缺乏准确、全面的材料科学数据，数据的应用将出现灾难性事件。另外，智能制造以制造过程的数字化表达，即数字孪生，及数据互联互通系统来实现全流程智能化控制，智能制造[15]以及“黑灯工厂”的建设不仅需要材料生产工序的数字化以及各工序间的信息传递，而且需要基于过程数据进行工艺参数的优化调控，因此，一套标准化的数据体系是材料智能制造的基础设施。

我国拥有的材料生产方面的国家标准众多，这些标准往往是针对一类材料甚至是某一型号材料的生产。我国材料数据标准的建设工作可以追溯到 2009 年材料科学数据共享网建设时期近 30 项标准草案的起草，内容涉及

材料数据库的建设、管理、数据采集等方面。然而由于当时尚无材料数据标准申报途径，材料数据标准化的进程一度停止。

目前设立了 10 余项特定材料的数据的相关国家标准，如金属材料的疲劳试验循环计数和相关数据缩减方法、塑料的可比单点数据的获得和表示等，但这些标准普遍存在全面性、通用性和互操作性的局限，使其应用效果及影响力受限，更是难以应用到数据库的建设与管理的指导上。材料领域学科众多，差异较大，同时材料数据来源于实验和计算，以及数据挖掘与机器学习获得的衍生数据，导致采集入库的数据的差异更大。而材料领域对数据的大规模的认识和学习是近些年才开始的，2014 年 FAIR 原则(Findable，Accessible，Interoperable，Reusable，简称 FAIR)作为科学数据管理与共享的普遍原则被提出[16]，成为材料数据及管理的评价规范，同时也是目前建立材料数据标准时被广泛接受和认可的行为准则。于是，基于此制定的材料数据统一描述标准成为材料数据采集与入库的理论指导基础。

北京科技大学正在牵头起草材料数据统一描述模型国家标准。该标准基于材料五要素，即成分、工艺、组织、性能及服役，通过它们的内在本质将材料数据关联在一起，形成了具有普适性的材料数据的描述准则。其中实验数据包括三大部分：

(1)成分数据：成分数据包括材料牌号/名称以及材料成分信息。前者是指该材料在其使用领域内的通用牌号或名称，对于新的材料可以自定义；后者包含组成材料的全部化学元素及其含量。

(2)工艺数据：工艺数据不仅包括原材料数据，还包括制备/合成/生产工艺数据。前者包括原材料性能与原材料生产方法，后者包括每一个工艺的名称以及该工艺涉及的所有工艺参数。

(3)性能数据：性能数据不仅包括材料性能及相关图片和视频，还包括取样信息及测试条件数据等。其中材料性能可以表示为数值型、字符型、图片、视频等类型，同时包含此性能所对应的显微组织照片和视频等；取样信息包括被检材料的来源、加工状态，样品在被检材料上的取样部位信息，以及样品的尺寸、形状信息；测试条件数据包括检测设备参数及型号，测试标准以及测试实验参数。

对于计算获取的数据的标准，包括计算任务名称、计算条件及计算性能数据。其中计算任务名称包含材料牌号、材料成分及计算内容名称；计算条件包含计算软件或公式信息、输入条件参数，以及对计算获得的直接结果进行分析的方法、分析计算条件及对数据的处理方法等；计算性能数据包含计算获得的中间结果和最终结果，以及经过分析处理后的计算分析数据。

对材料性能属性也进行了统计分类，形成了包括力学、物理、化学、电学、磁学、光学、热学、电化学、生物学、环境、晶体结构、工艺性能在内的 12 大类材料性能的数据字典。目前已收集的属性名称达到 780 余个，尚有相当数量的属性将在今后进一步收集。

材料数据统一描述模型标准作为材料数据描述的通用标准，在面对不同专业、材料类别、数据来源时需要进一步细化粒度。近几年随着团体标准的出现以及中关村材料试验技术联盟团体材料基因工程标准委员会的建立，材料数据团体标准的建设也迅速发展起来。2019 年，中关村材料试验技术联盟团体材料基因工程标准委员会通过了团体标准《材料基因工程数据通则》(下简称“通则”)，成为目前材料基因工程领域的第一项对材料数据标准体系进行约束的材料数据团体标准。该标准通则提出的按照数据的作用及来源进行分类的方法，即元数据、源数据及衍生数据三类数据，成为了材料数据团体标准的制订准则。根据通则的术语定义，元数据是指与样品和数据的条件有关的数据，结合前述讨论，材料数据的元数据具体指的是实验数据中的制备/合成/生产工艺数据、取样信息及测试条件数据等，以及计算数据中的计算条件。源数据是指测量或计算产生的原始数据，而衍生数据，通则中定义为对原始数据进行加工或分析后产生的结果数据。除了第一性原理计算及热力学计算获得的高通量计算数据，以及高通量实验及高通量表征获得的数据外，其他来源的数据量都相对较小，属于“小数据”，但无论是高通量数据还是“小数据”，其价值都很高。而对于通过高通量技术获得的数据，往往基于这些数据进行机器学习等数据分析获得描述材料特征的衍生数据，其价值通常会高于单一数据。当然，基于不同的分析处理技术和算法获得的结果不同，价值也会有差异。

作者团队近几年来起草了高温合金、轻金属、陶瓷基复合材料、催化材料、稀土材料、能源材料等材料的数据标准，以及热力学、相场动力学计算数据的标准等数据标准草案，并向中关村材料试验技术联盟团体材料基因工程标准委员会提交了高温合金数据、稀土合金数据、材料元数据、计算数据管理标准规范、热力学计算数据、相场动力学计算数据等多项材料数据团体标准，以期材料数据细粒度的团体标准能够尽快满足各材料学科的需求。

4 数据采集技术

由于材料学科方向的多样性以及材料五要素的复杂

性及关联性，几十年来高质量数据基本都是依赖具有丰富的材料领域知识的研究人员人工采集。近年来，随着文本挖掘技术的发展，国内外学者开始尝试采用文本挖掘技术从文献中抽取除图表等以外的信息来补充材料文献数据，如 Olivetti 等[17]对自然语言处理过程、方法及在数据驱动的材料科学研究中的应用进行了全面系统的综述，Court 等[18, 19]将自然语言处理技术与机器学习相结合自动构建了磁性材料及超导材料的性能数据库及相图，并具备预测相变温度的能力。文本挖掘技术在高温合金文献数据提取上也取得了突破[20]，充分证明了文本挖掘技术与材料数据自动提取技术在材料数据获取上的潜力，并在准确性与通用性上达到平衡，今后将成为材料数据研究的热点之一。

北京科技大学陈宁教授提出了利用互联网资源的高效数据智能收集方案，即基于互联网信息资源搜索，创新性地提出了一种集成数据收集、分析整理、智能审核和统计评估四大功能模块的材料数据处理流程，如图 3 所示，实现了智能化和高效率的数据整理和分析校验的完整功能。并利用该核心技术，从 Web of Science 上 20 万条文献的摘要中获得有效性能数据 10 464 条，完成了“锂离子电池”和“固体氧化物燃料电池”等能源材料专用数据库，并自动汇交于 MGEdata. cn 平台。

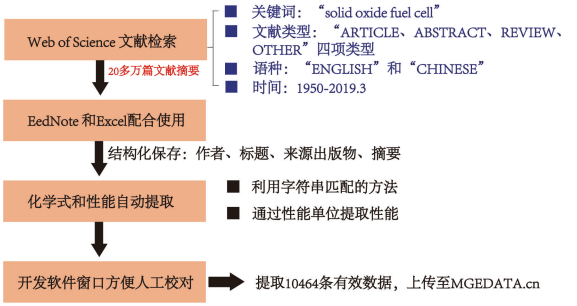

图 3　以固体氧化物燃料电池材料数据搜索为例，实现文献摘要数据的下载整理及材料化学式、结构参数及性能数据的智能收集的流程图

Fig. 3　Taking the materials data collection for the solid oxide fuel cell as an example, the flow chart shows downloading and sorting of literature abstract data, and intelligent collection of chemical formula, structure and performance data

值得指出的是，尽管文本挖掘技术发展迅速，但由于材料学科的多样性与复杂性，该技术在材料领域的全面应用还需较长的时间以及人工的辅助，目前从文献获取信息全面且准确的高质量材料数据仍需要人工采集。

5　数据存储的动态容器技术

结构化数据库是材料研发人员最易理解的数据库建设方法，但应对各类材料不同类型和结构的数据，以及不同材料的多种不同性能属性的存储需求，一个统一的结构化数据库是远不能满足通用性存储的要求的。而建设国家级大型材料数据库是近几年各国应对大数据时代下材料科学与技术发展的举措。国家级材料基因工程数据平台面临的挑战之一就是要将不同尺度的计算数据和文献收集的实验数据按照材料数据标准草案的要求进行大量信息的存储，并支持高通量的计算结果数据及高通量实验或表征数据的存储。

广泛用于云计算与服务平台虚拟化的容器技术具有轻量化、独立化的特点[21-23]。Liu 等[24]提出用户友好的动态容器技术，用户自定义数据存储结构，突破了各类材料数据存储的多样性问题，实现材料数据的无模式存储，如图 4 所示，使得存储的每条数据可以有多种属性，属性的类型可以是数字、字符串、日期等基本数据类型，也可以是数组或子文档等，实现了数据存储形式的普适性，即数据模型的逆规范化(denormalizing)，这样既解决了计算、实验、表征数据的多元异构问题，以及材料数据标准对数据内容要求的鲁棒性与数据存储的灵活性的矛盾，也同时提高了查询的速度。

6　材料数据的知识产权保护

通常提及的知识产权保护往往指的是专利[26, 27]，但近年来我国对数据产权的关注及研究逐步深入，紧跟大数据时代特征。黄海瑛等[28]对我国数据产权相关的政策进行了总结分析，认为 2021 年进入大幅增长期，呈现稳中有增的态势，产权的保护政策包括通知、意见、建议、纲要及条例等主要类型，其目的在于解决数据保护、应用与交易、制度与标准规范等相关问题。Cao[29]提出采用机器学习算法对教育数据知识产权的表达进行分类并赋予不同的保护措施。而国际上学者们对数据的知识产权问题在 21 世纪初就已经相当重视，2011 年发表在 *Science* 杂志的关于干细胞数据管理的文章不仅提及知识产权，而且将个人与权利保护及科学发展关联在一起[30]。

科学数据是以数据形式表达的知识，由于其内涵的科学性，使得每一条数据都具有很高的价值，因此，科学数据的知识产权保护与数据共享既矛盾又统一，是大数据发展的关键问题。只有对科学数据进行产权保护，才能使数据生产者有意愿将数据与他人共享，才能保证数据来源，否则数据共享终将面临无数据可共享的尴尬局面[31]。中国科学院地理科学与资源研究所刘闯研究员[32]2010 年在国内率先将数据对象标识符(digital object identifiers，DOI)系统用于地理科学数据，使数据具有了全球唯一的“身份证”，引领了基于科学资源标识的数据资源知识产权保护的发展。

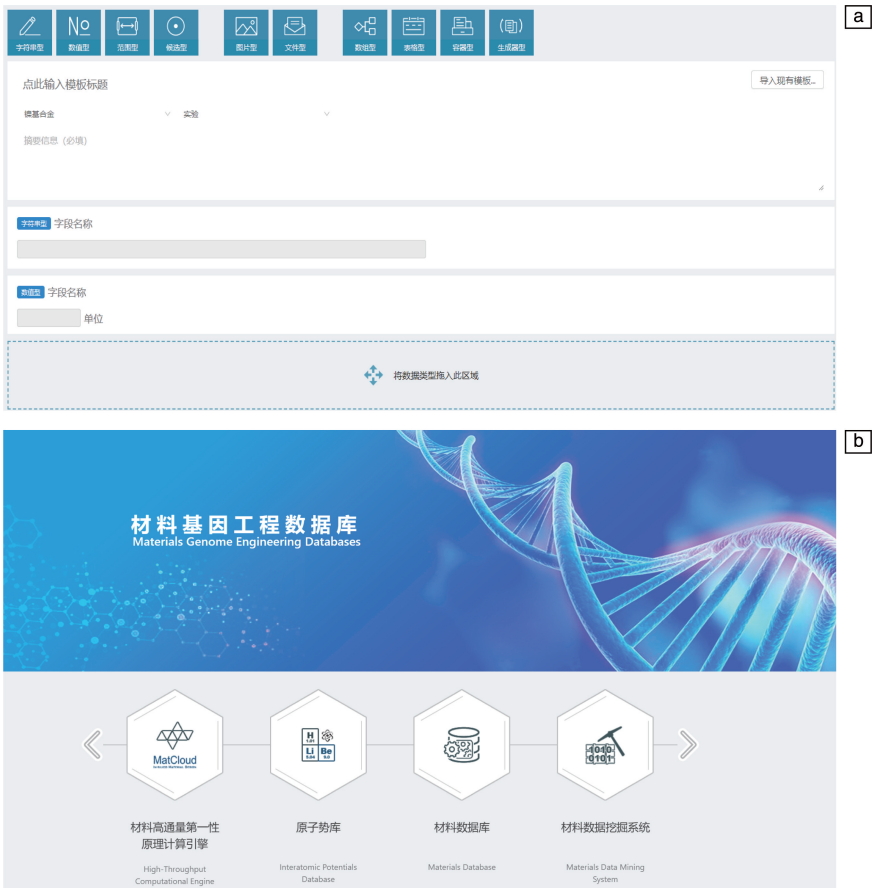

图 4 在材料基因工程数据库平台 www. mgedata. cn 上基于动态容器技术的数据存储，其中根据材料数据的特点将数据类型分成字符串、数值、图片、文件、数组、表格、容器、生成器以及范围、候选 10 个类型，用户可用拖拽形式增加新的数据类型(a)[24]；材料基因工程数据库平台 www. mgedata. cn 主页，平台已上线并实现了数据的存储、管理及数据资源深度开发利用(b)[25]

Fig. 4 Data storage in the MGE database platform www. mgedata. cn, according to the characteristics of materials science, the data are classified into 10 types including strings, values, pictures, files, arrays, tables, container, generator, range and candidate, and users can easily add new data types by dragging (a)[24]; the homepage of the MGE database platform www. mgedata. cn, which has been launched online, and storage, curation and in-depth utilization of data resources have been realized (b)[25]

材料科学数据知识产权保护从 2014 年提出，也采用了 DOI 系统标识材料数据，依据国家标准《GB/T 32843-2016 科技资源标识》形成了如图 5 所示的材料数据标识形式。材料数据 DOI 包括数据拥有者所在单位代码、材料数据分类代码、产生数据的资助项目序号、DOI 注册时间、数据库/数据集序号以及序列号等部分。其中材料数据分类代码以 mater 开头，可以清晰地从 DOI 信息中分辨出来，具有高度的可辨识性，后面的 4 位数字依据图 1 的材料数据分类方法中的二级和三级的名称代码确定。因此，材料数据 DOI 包含的信息可以作为一类元数据，成为数据管理的重要参数。目前材料科学数据的 DOI 系统已经用于 MGE 数据平台的材料数据产权保护，系统可以自动生成数据 DOI 的序列号，使数据生产者的知识产权保护可一键式生成。

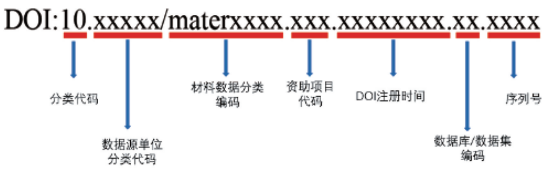

图 5 用于材料科学数据知识产权保护的 DOI 系统

Fig. 5 DOI system for the intellectual property protection of material science data

7 材料数据的质量管控

数据质量的重要性在大数据时代显得尤为突出。随着社会经济的飞速发展，数据资源的数量也在激增，由于技术、人员等方面壁垒的存在，数据的科学性、系统性、完整性、有效性等方面成为科学数据工作者亟待解

决的问题。

对材料基因工程数据库的数据质量目前可以从两方面开展相应的管控，其一是数据满足 FAIR 原则的管控，其二是通过数据挖掘与机器学习技术对数据质量可靠性进行评估。

FAIR 原则旨在推进材料科学数据可发现、可获取、可互操作、可再利用[33]。该原则也成为材料数据共享的基础原则，本文第三章的材料科学数据标准系统就是以国家标准及团体标准的方式，对材料数据的内容进行描述和约束，在数据平台中以数据模板的形式呈现，使数据生产者能够以标准化的形式采集数据，同时数据审核者对采集的数据进一步把关。数据在入库时会经过两次质量审查，一方面平台维护人员将对有格式问题的异常数据进行检查，发现因输入错误造成的数值明显不符合常规的错误，如数量级的差异等；另一方面，平台专家对入库数据进行抽查，从而通过标准的技术屏障和专家的专业屏障对数据质量形成双重保证。

将数据挖掘算法用于材料数据旨在从数据中寻找规律，并依据相应的数据规律来发现异常点或离群点。材料数据的“小数据”，区别于大数据的其中一个功能是发现不寻常的数据，离群点是材料创新的来源之一，分析该数据有可能获得创新性的知识。因此，通过分析离群点来判断材料数据质量仅能作为数据质量评估的手段之一，实际使用时需要慎重。而数据挖掘算法也是缺失数据补齐的重要手段之一，主要针对早期数据可能存在的一些数据点缺失或文献中未披露而目前难以依据数据获取路径再次获得的情况。目前以提升缺失数据补齐精度为目标的适用算法仍在探索和优化中。

8　材料基因工程数据生态系统的构建

材料基因工程作为材料创新研究的方法论，正处于方兴未艾之时。对其中发展空间最大的要素——数据，应构建材料基因工程数据生态系统，使材料数据的发展形成一个完整的闭环，夯实材料数据未来产业发展的理论基础。基于材料科学数据共享网建设的经验，在“十三五”期间，逐步在数据分类、数据标准、数据采集、存储、共享、数据质量评估及数据库建设七大关键方面的研究及其关键技术上取得了较大的突破，使材料基因工程大数据及数据库建设形成了科学、系统、严格的技术闭环，建成了材料基因工程数据生态系统，如图 6 所示。相较于材料数据全生命周期特征的技术闭环，由数据生产者、数据用户及市场构成的社会环境逐步孕育，其完善程度还存在差异。相较于技术上的闭环，由数据生产者、数据用户及市场构成的社会环境逐步孕育。国外部分数据库，如 Pauling file、Atomwork 等，因其成熟的建设历程及稳定的数据生产者队伍，形成了一定数量的市场，虽然规模不大，但其引领作用不可小觑。

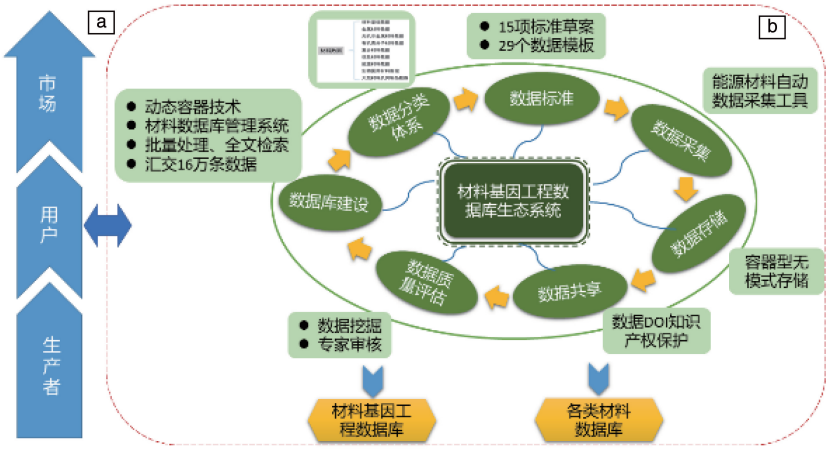

图 6　材料基因工程数据生态系统的关键特征及部分代表性技术成果：(a) 由生产者、用户及市场构成的数据生态系统的社会环境，(b) 材料数据全生命周期特征的技术闭环

Fig. 6　The key elements and the obtained research results in the data ecosystem of Materials Genome Engineering: (a) social environment comprised of data producers, users and markets; (b) technical closed loop

9　结　语

在大数据时代的大背景下，计算机和软件技术为材料加速创新和智能制造提供了底层的技术基础，基于数据驱动的材料研究方法迎来重大发展机遇。材料数据具有体系多、数据量小、维度高的小数据特点，显示出与其他领域科学数据的显著差异。数据目前是材料基因工程的关键要素，本文提出的材料基因工程数据生态系统

基本涵盖了材料数据理论基础涉及的数据分类体系及数据标准，数据库建设的采集、存储等基本步骤，以及材料数据管理的知识产权保护及数据质量控制等，形成了材料基因工程数据生态系统的基础架构，并取得了较好的成果。近几年，在国家项目的支持下也培育出一些具有典型材料基因工程特色的数据库或数据平台，如北京航空航天大学孙志梅教授团队开发的加速新材料研发与设计的可视化高通量自动流程计算和数据管理智能平台 ALKEMIE[34, 35]，集成了高通量计算和支持高通量计算结果及文献数据存储的数据库，以及具有成分筛选与性能预测功能的数据挖掘工具，打造了计算与数据存储分析相交互的新材料研发平台。中国科学院物理研究所刘淼博士等开发的第一性原理计算数据库 Atomly[36]，在 Materials Project 数据库理念基础上进一步发展其功能与覆盖范围，与 ALKEMIE 相同，两者在材料数据获取手段、数据量和精度上显示出鲜明的特色和不断扩展服务的功能。随着各领域的数据生态系统的不断完善和发展，以及材料基因工程理念被更多人学习、接受并应用在不同材料体系研究中，材料基因工程数据生态系统也将进一步得到优化。

材料基因工程数据生态系统未来将在以下几个方面得到深入的研究和发展：

(1)数据标准

数据标准是数据生态系统的一个要素，数据标准的内容则全方位覆盖数据生态系统。数据标准体系的构建与标准的制定决定了材料基因工程数据生态系统的发展。尽管材料基因工程数据标准体系尚未健全，但在今后，涉及材料数据的采集、分析加工、交换与接口、存储与治理[37]、垃圾数据处理、开放共享及服务、数据安全、产权与隐私保护等数据全生命周期中的问题，将成为材料基因工程数据标准体系的基本框架，并将随着材料基因工程研究的深入，逐步形成相关的标准。

在科学数据标准申请方面，全国科技平台标准化技术委员会(SAC/TC486)主要负责科学数据相关国家标准的申报立项工作，目前已申报立项的国家标准达到几十项，主要涉及科学数据共性问题，而关于学科领域数据标准的研讨刚刚起步。另外，国家对团体标准建设的支持使标准申请的周期大大缩短，在中国材料与试验团体标准(Chinese Standards for Testing and Materials，CSTM)委员会下设立了材料基因工程领域委员会(FC97)，使材料基因工程相关团体标准能够迅速落地，目前随着团体标准制定工作流程的不断完善，材料数据团体标准申请已呈现出稳步上升趋势，相信今后会有相当数量的团体标准在此立项。

(2)数据来源

国家创新发展对材料的需求推动材料基因工程不断发展，材料数据的来源也将有较明显的变化。高通量计算，尤其是高通量第一性原理计算[38, 39]及相关软件工具与平台的推广，以及高通量热力学计算的应用[40]，在新材料发现和数据挖掘中将显示越来越重要的地位。高通量实验，包括典型的组合芯片实验以及在各实验室陆续开展的小型高通量实验，与高通量表征，包括正在投入的北京怀柔的同步辐射光源与广东东莞的散裂中子源等大型科学装置的应用，以及以中国科学院物理研究所金魁教授为代表自行开发的研究装置[41]，成为高通量实验和高通量表征数据的重要数据来源，这方面美国等发达国家发展得更好[42]。而如何定义一条数据的问题也在数据管理上显得愈发突出。

今后针对高通量计算、实验及表征数据的标准将会得到较大发展，这是由于这类数据流程性强，标准相对容易推广，同时相关数据标准将更显著地呈现数字化使用的特点，即可机读(ISO smart)。

除高通量计算与表征外，图像处理，尤其是微观组织图像分析处理，成为近年来的材料基因工程及材料组织定量化表征的研究方向之一，且热度不断升温。随着对材料性能预测和优化的相关研究深入，通过解析组织特征参量，以及通过卷积神经网络等方法由微观组织直接获得材料性能的研究将成为未来研究热点之一[43]，并将成为重要的数据来源。

(3)数据的价值

材料数据的价值不仅在开放共享过程中显示，随着全社会更大范围的数据质量提升与数据价值评估体系的发展和完善，材料数据的价值将进一步展示，将不仅服务于材料基因工程相关材料研究，而且在数字时代智能制造等变革性科学生产实践及高等教育人才培养过程中发挥更大的效力。随着材料基因工程数据生态系统的架构不断完善，数据生态系统将为未来材料数据进入大数据市场和参与数据交易活动奠定理论和应用基础。届时，数据安全、知识产权保护以及隐私保护等方面问题及其解决方案的研究与应用推广也将会成为焦点问题之一。

参考文献 References

[1] DAVIES T. Oryx[J], 2011, 45(2): 304-305.

[2] DING L, LEBO T, ERICKSON J S, *et al*. Journal of Web Semantics [J], 2011, 9(3): 325-333.

[3] OLIVEIRA M I S, de FÁTIMA BARROS LIMA G, LÓSCIO B F. Knowledge and Information Systems[J], 2018, 61: 589-630.

[4] AAEN J, NIELSEN J A, CARUGATI A. European Journal of Information Systems[J], 2022, 31(3): 1-25.

[5] ANWAR M J, GILL A Q, HUSSAIN F K, *et al*. EURASIP Journal on Wireless Communications and Networking[J], 2021, 2021: 130.

[6] HAYASHI T, ISHIMURA G, OHSAWA Y. IEEE Access[J], 2021, 9: 52266-52276.

[7] ALLISON J, COMPANY F M. National Research Council. Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security[C]// 2010 AIChE Annual Meeting. Washington D.C.: The National Academies Press, 2010.

[8] WARD C. Materials Genome Initiative for Global Competitiveness[C]// 23rd Aeromat Conference & Exposition. American Society for Metals, 2012.

[9] LIU Z K. Engineering[J], 2020, 6: 604-608.

[10] 尹海清, 姜雪, 张瑞杰, 等. 中国材料进展[J], 2017, 36(6): 401-404.
YIN H Q, JIANG X, ZHANG R J, *et al*. Materials China[J], 2017, 36(6): 401-404.

[11] 尹海清, 徐斌, 姜雪, 等. 中国科学数据[J], 2019, 4(1): 136-145.
YIN H Q, XU B, JIANG X, *et al*. China Scientific Data[J], 2019, 4(1): 136-145.

[12] YIN H Q, JIANG X, LIU G Q, *et al*. Chinese Physics B[J], 2018, 27(11): 124-129.

[13] 尹海清, 刘国权, 姜雪, 等. 科技导报[J], 2015, 33(10): 50-59.
YIN H Q, LIU G Q, JIANG X, *et al*. Science & Technology Review[J], 2015, 33(10): 50-59.

[14] 国家材料科学数据共享网[EB/OL]. (2001-12-19)[2021-12-01]. http://www.materdata.cn/.
Materials Scientific Data Sharing Network[EB/OL]. (2001-12-19)[2021-12-01]. http://www.materdata.cn/.

[15] ZHOU C Y, CAO Q L. The Journal of Networks Software Tools and Applications[J], 2019, 22: S8647-S8655.

[16] WILKINSON M, DUMONTIER M, AALBERSBERG I, *et al*. Scientific Data[J], 2016, 3: 160018.

[17] OLIVETTI E A, COLE J M, KIM E, *et al*. Applied Physics Reviews[J], 2020, 7(4): 041317.

[18] COURT C J, COLE J M. Scientific Data[J], 2018, 5: 180111-180123.

[19] COURT C J, COLE J M. npj Computational Materials[J], 2020, 18: 1-9.

[20] SuperalloyDigger[EB/OL]. http://superalloydigger.mgedata.cn/

[21] 何立民. 单片机与嵌入式系统应用[J], 2021, 21(6): 4-6.
HE L M. Microcontrollers & Embedded Systems[J], 2021, 21(6): 4-6.

[22] PAHL C, BROGI A, SOLDANI J, *et al*. IEEE Transactions on Cloud Computing[J], 2019, 7(3): 677-692.

[23] 李佳曦. 信息通信技术[J], 2020, 14(6): 26-32.
LI J X. Information and Communications Technologies[J], 2020, 14(6): 26-32.

[24] LIU S L, SU Y J, YIN H Q, *et al*. npj Computational Materials[J], 2021, 88: 1-8.

[25] 材料基因工程数据库[EB/OL]. https://www.mgedata.cn/storage/template/new.
Materials Genome Engineering Databases[EB/OL]. https://www.mgedata.cn/storage/template/new.

[26] SIRI J G S, FERNANDO C A N, de SILVA S N T. Recent Patents on Nanotechnology[J], 2020, 14(4): 307-327.

[27] 唐彬, 潘麟, 明成昆. 上海商业[J], 2021, 09: 180-181.
TANG B, PAN L, MING C K. Shanghai Business[J], 2021, 09: 180-181.

[28] 黄海瑛, 文禹衡. 图书馆论坛[J], 2022, 42(3): 18-30.
HUANG H Y, WEN Y H. Library Tribune[J], 2022, 42(3): 18-30.

[29] CAO J W. Complexity[J], 2021, 2021: 1909518.

[30] MATHEWS D J H, GRAFF G D, SAHA K, *et al*. Science[J], 2011, 331(6018): 725-727.

[31] VIRGINIA G. Nature[J], 2016, 529(7584): 117-119.

[32] 刘闯. 全球变化数据学报[J], 2020, 4(2): 101-109.
LIU C. Journal of Global Change Data & Discovery[J], 2020, 4(2): 101-109.

[33] 宋佳, 温亮明, 李洋. 情报资料工作[J], 2021, 42(1): 57-68.
SONG J, WEN L M, LI Y. Information and Documentation Services[J], 2021, 42(1): 57-68.

[34] WANG G J, PENG L Y, LI K Q, *et al*. Computational Materials Science[J], 2021, 186: 110064.

[35] 孙志梅. 数据驱动的材料跨尺度建模与设计[C]// 北京高精尖论坛--第一届青年材料科学家论坛文集. 北京: 北京材料基因工程高精尖创新中心, 2020.
SUN Z M. Data-Driven Across Scales Modeling and Design for Materials[C]// Proceedings of the First Forum for Young Materials Scientists of Beijing Advanced Technology Forum. Beijing: Beijing Advanced Innovation Center for MATERIALS GENOME ENGINEERING, 2020.

[36] Atomly 材料科学数据库[EB/OL]. https://www.atomly.net/.
Atomly Modernize the Materials Science[EB/OL]. https://www.atomly.net/.

[37] GROSSMAN R L. Cancer Journal[J], 2018, 24(3): 122-126.

[38] JAIN A, HAUTIER G, MOORE C J, *et al*. Computational Materials Science[J], 2011, 50(8): 2295-2310.

[39] GARRITY K F, BENNETT J W, RABE K M, *et al*. Computational Materials Science[J], 2014, 81: 446-452.

[40] ZHANG C, JIANG X, ZHANG R J, *et al*. Computational Materials Science[J], 2019, 167: 19-24.

[41] YUAN J, CHEN Q H, JIANG K, *et al*. Nature[J], 2022, 602: 431-436.

[42] FENG R, ZHANG C, GAO M C, *et al*. Nature Communications[J], 2021, 12: 4329.

[43] FEI Y, WANG K C P, ZHANG A A, *et al*. IEEE Transactions on Intelligent Transportation Systems[J], 2020, 21(1): 273-284.

(编辑　惠　琼)