科学通报 2024 年 第 69 卷 第 9 期: 1164 ~ 1174
《中国科学》杂志社 SCIENCE CHINA PRESS
香山科学会议专栏 观 点
第S69次学术讨论会·科学数据治理与利用的前沿和热点

凝聚态物质科学科研数据管理与应用

王丹[1,2,4†], 周明波[1,2†], 黄东宸[1,2], 李云龙[3], 林泽丰[1], 刘俊德[1], 朱天念[1,2], 竺云[4], 李明星[1,2], 肖睿娟[1,2], 袁洁[1,2*], 翁红明[1,2*]

1. 中国科学院物理研究所, 北京凝聚态国家实验室, 北京 100190;
2. 中国科学院凝聚态物质科学数据中心, 北京 100190;
3. 中国科学院前沿科学与教育局, 北京 100864;
4. 天津师范大学物理与材料科学学院, 天津 300387
† 同等贡献
* 联系人, E-mail: yuanjie@iphy.ac.cn; hmweng@iphy.ac.cn

大数据已经以不可阻挡的脚步踏入了社会生活, 大到国家范围流行病趋势的准确预测[1]、纳斯达克上市公司市值估算(http://busmiamiedu/umbfc/common/files/papers/Karabulutpdf); 小到机票最佳购买时机的选择[2]、出行自动导航等[3]. 作为一种新的技术, 大数据已对人们的生活、生产带来了深远的变革性影响[4,5].

科学研究同样如此, 以天文学为例, 2000年斯隆数字巡天(Sloan Digital Sky Survey)项目启动的时候, 位于新墨西哥州的望远镜在短短几周内收集到的数据, 已经比天文学历史上总共收集的数据还要多[6]. 2016年在智利投入使用的大型视场全景巡天望远镜(Large Synoptic Survey Telescope)在5 d之内就获得高达1.4×2^{42}字节的信息[7]. 天文学领域的变化在其他领域也在发生. 如2003年, 人类第一次破译人体基因密码时, 10年完成了30亿对碱基对的排序[8]. 大约10年之后, 世界范围内的基因仪每15 min就可以完成同样量的工作[9,10]. 另外, 得益于大规模计算能力、存储能力和科学仪器的共享支持, 大型强子对撞机(Large Hadron Collider, LHC)每年产生的20 PB(1 PB=1024 TB)数据被全球联网的上百万个节点进行存储和处理[11]. 借助这种方式科学家抓住了万亿分之一的几率, 于2012年7月4日宣布发现希格斯粒子[12]. 逐渐地, 在科研领域诞生了实验、理论和仿真之外一种新的研究范式(图1)——数据密集型科学发现[13~15](data-intensive scientific discovery).

大数据技术不仅正在加速科学发现, 而且从根本上改变了人类探索未知世界的思维和行为方式[17,18]. 譬如当初开普勒在经典力学框架还没有建立的情况下, 通过分析其老师第谷·布拉赫长期积累的天文数据就得到了描述行星运动的三大定律[19]. 基于大数据的全面性, 人们可以跳过建模、理论和推演的过程来直接获得事物之间的关联, 这使得复杂体系的内在关系与演化规律不再是遥不可及的研究目标.

物质科学的研究对象是非常复杂的体系, 其涉及粒子的



王丹 中国科学院物理研究所与天津师范大学物理与材料科学学院联合培养硕士研究生. 主要研究方向为机器学习、数据库开发等.



袁洁 博士, 主任工程师, 兼任中国科学院凝聚态物质科学数据中心副主任. 从事科研大数据规范化探索及相关仪器部件的研制.



翁红明 博士, 研究员, 兼任中国科学院凝聚态物质科学数据中心主任. 主要研究方向为拓扑量子态及其材料计算、磁光效应、非线性光学效应等第一性原理计算.

© 2024《中国科学》杂志社

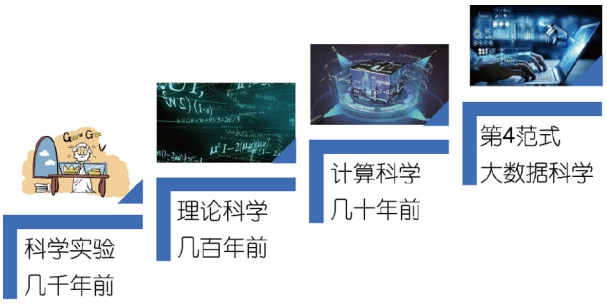


图 1 科学研究范式的发展[16]
Figure 1 The development of scientific research paradigm[16]

数目巨大(10^7~10^{21})，粒子的空间排列形式复杂，粒子间还存在着各种相互作用和关联. 试图完整地描述这样一个系统已经超越了以“The Theory of Everything”为起点的还原论思想的极限[20]. 高通量计算[21]、高通量实验[22]和人工智能技术[23]的发展，进一步扩展了物质研究的空间，特别是对包含多种元素的材料体系. 由于多体相互作用引入的复杂性，探索这类材料的组合空间，对传统的科学研究范式提出了巨大的挑战. 大数据技术抛开直接追求因果关系，换而追寻相关关系的思路为物质科学的发展打开了一条新的道路[24]. 事实上，近几年在信息科技快速发展的背景下，以大数据为基础的人工智能等手段已经开始与凝聚态物理和材料科学等领域的研究方法相结合，将理论、计算和实验测量等数据融会贯通，在加速材料发现、优化制备方案、创新器件应用等方面显示出独特的优势和潜力[25]. 逐渐发展出了材料信息学这门新兴的学科[26]. 随机森林和神经网络等算法已成为材料领域常用的数据处理和分析工具，在性能[27]、相图[28]和结构[29]的预测方面取得了显著的效果. 但当大数据为凝聚态物质科学、材料科学等研究打开了更全面、更深入的视角和思路同时，如何实现数据的规范化治理与高效应用也成为越来越重要和亟须解决的问题[30].

1 数据规范化

1.1 数据规范化意义

数据密集型新范式的出现标志着假设驱动的传统研究(“先假设后实验”)向“先实验后假设”的发现模式转变[31]. 因此，数据成为当前科学发现的先决要素，是物质研究新范式的基础[32]. 面对来自不同课题组数以万计的科学实验，存在数据收集量大和数据标准难以统一等难题. 一方面，实验数据的分析仍仅限于实验人员个人进行，甚至同一实验室成员之间都无法看懂对方的实验数据，相同研究领域的研究人员也会因数据隔离而重复相同的工作；另一方面，大量原始数据和“失败数据”仅停留在实验室里或者被抛弃，未能获得充分的利用. 学科之间的数据传递存在着较高的壁垒，使得交叉合作的研究难以开展. 对于研究者来说，从大量繁杂的数据中提取出想要的信息无异于大海捞针. 种种因素都导致了资源的浪费和科研发展的滞缓. 以凝聚态物质科学为例，其数据来源众多，分类繁杂，格式多样，分布零散，重复不一，质量不齐. 亟需通过有效的数据治理，形成高质量的数据资产[33]. 这也是在凝聚态物质科学领域发展推广“大数据+人工智能”研究范式的前提条件.

1.2 数据规范化发展趋势和挑战

数据规范化管理已成为国内外发展的大趋势. 近些年，随着数据量快速增长，我国出台了一系列数据标准规范[34]. 国际上，数据治理研究所(Data Governance Institute, DGI)总结了数据治理的10大关键要素，提出了DGI数据治理框架. 国际数据管理协会(Data Management Association, DAMA)也颁布了《数据管理知识体系指南-DMBOK》[35].

许多科研单位意识到数据的价值，开始在科研数据的管理上开展相应的工作. 例如，德国波鸿鲁尔大学的Ludwig课题组[36]开发了名为Research Data Management的数据管理平台. 他们以商用的文件管理软件为基础，结合多个自主编写的数据处理程序，形成了交互式的全流程实验数据管理模式. 在该平台下，数据管理分为两种层级：一种严格按照既定的实验流程进行，一旦产生原始数据，将被程序自动捕获分析，形成最终的实验结果；另一种仅记录实验条件，数据分析具有更高的自由度，更适合专业的数据处理软件，如透射电子显微镜的数据. 其他的高通量实验数据管理平台也有类似特点，如MEAD[37]、HTEM[38]和Materials Atlas Project[39]. 无机晶体结构数据库(Inorganic Crystal Structure Database, ICSD)是由德国FIZ Karlsruhe提供的世界上最大的无机晶体结构数据库. Park等人[40]利用ICSD(https://icsd.products.fiz-karlsruhe.de)计算出15万个X射线衍射(X-ray diffraction, XRD)图谱，然后利用卷积神经网络(convolutional neural networks, CNN)模型从模拟的XRD图谱中预测出结构信息. Zaloga等人[41]利用ICSD数据库实现了对晶体的分类.

在物质科学科研领域，数据管理虽已有一定的进展，但目前的数据库仍少有以实验数据为主的，原因之一是计算数据的形式容易结构化和统一，而物质科学实验的仪器种类繁多，实验条件各异，这种复杂性导致在原始数据的基础上需要进一步的描述，从而增加了数据规范化的难度. 如，即使ICSD已产生巨大的科研价值，但其数据库的每个空间组(标签)表示的结构数量也存在很大差异，规范化工作也是巨大的挑战[42]. 此外，对于科研数据，质量保证或评估往往难以实施，这也给实验数据的收集和标准化带来了困难. 流程固定的实验过程所产生数据的管理，对于数据管理平台的要求较为明确，开发者只需要串联几个固定的仪器设备，制定一定的操作规范即可，这可能是化学、生物等领域最早大规模使用材料基因工程的原因之一. 然而，对于凝聚态物质研究而言，研究涉及的过程复杂、仪器设备种类繁多、所产生的数

据形式多样. 不同实验过程用到的硬件设备和数据处理软件相互间的兼容性较差. 而且大多数课题组的核心竞争力主要集中在新材料的开发和研究, 在信息技术和软件开发上的投入有限, 难以串联各个实验, 导致在物质科学研究领域, 数据管理的尝试通常也因为缺少合适的信息化工具而难以实施.

不仅如此, 在实际科研过程中, 数据规范化还面临着人员习惯的挑战. 实验人员已形成固有的记录方式, 纸质版实验记录仍为主流. 为了便于数据交换和数据共享, 科研人员需要对数据有一个共同的认识和理解——数据规范化是摆在我们面前的一项亟须实施的任务. 科研人员应主动适应新型数据管理方式, 打破固有习惯[43].

1.3 数据规范化的实现

数据规范化需要全面考虑, 确立符合FAIR原则[44](findability, 可发现; accessibility, 可访问; interoperability, 可互操作; reusability, 可重用)的数据标准, 从根本打破数据壁垒, 消除“数据孤岛”, 从而提高数据资源的利用效率.

(1) 制定数据标准. 包括但不限于元数据、主数据、数据标签、数据模型等, 涵盖数据从采集、存储、处理、应用、共享、归档到销毁的整个生命周期, 形成完整链条.

(2) 建立数据管理平台. 数据规范化过程会遇到数据如何存储、高效处理和评估等的问题. 合适的数据管理平台可实现数据整个链条的管控. 同时, 数据管理平台应覆盖多种数据类型、数据全生命周期, 可以确保数据的安全性、共享性. 直到2021年, 已经有Labfolder[45]、eNovator[46]、LabArchives[45]和eLabFTW[47]在内的30多种软件工具被开发用于实验室数据的收集和管理. 中国科学院凝聚态物质科学数据中心根据学科特点, 抓住数据流转过程中的难点和痛点, 开发了物质科学电子实验室(Electronic Laboratory for Material Science, 以下简称“MatElab”), 覆盖科研数据的全生命周期.

(3) 优化数据流程、完善设备数据自动化采集. 人为记录数据总会受到环境、主观意识等因素的限制, 数据精确度、准确度、全面性都无法得到保证. 例如, 在脉冲激光沉积(pulsed laser deposition, PLD)薄膜制备中, 涉及的参数多达60多项, 凭借人工手动输入, 不仅耗费时间, 数据也会与实际产生偏差, 甚至出现错误数据. 为消除人为不利因素的影响, 需实现设备数据采集的自动化.

(4) 建立标准数据库. 数据库是有机整合各种相关信息的有效手段, 是数据系统化、结构化、标准化、可视化的集成展现平台. 在材料科学中, 材料对其制备过程很敏感, 同时结构表征、物性测量每一步都对最终结果产生影响, 但传统的材料数据库和科学文献通常没有提供足够的信息, 这给材料数据留下了很大的不确定性. 因此, 新数据库不仅需要收纳以材料信息为主的数据, 还需要同时跟踪实验和分析原始数据, 形成机器可读的材料描述符, 用于训练机器学习模型. 对于具有特定研究目标的物质科学研究, 科学数据的生成、处理和分析等各个流程需要更加规范. 对于自由探索式的研究, 也应尽可能多地保留数据痕迹, 增加数据的可读性和可重复性. 美国通用电气航空公司(GE Aviation)研究人员在进行低铼单晶合金涡轮叶片的设计之前, 基于长期积累的合金数据库, 通过神经网络模型进行处理, 预测了几百种合金的性能, 从而筛选出最合适的成分, 开发了两种新的镍基高温合金(牌号为N515和N500), 开发耗时由6年缩短为2年, 同时研发成本也大大降低[48]. 美国劳伦斯伯克利国家实验室利用组合芯片技术, 通过快速表征, 在一次实验中筛选出前人多年研究才发现的$Ba_2Sr_2CaCu_2O_{8+x}$和$YBa_2Cu_3O_{7-x}$两种超导材料[49].

(5) 结合技术实现数据价值最大化. 将快速发展的大数据、人工智能、云计算等技术应用到实际科研中, 减少实际数据处理、数据分析的工作, 提高科研效率, 为数据赋能. 在确保数据安全、数据保密的前提下, 促进数据开放共享, 提高数据重用率[50], 降低因重复研究造成的资源浪费.

基于上述的需求与考虑, 我们提出了一种治理凝聚态物质科学数据的方案[51], 其总体规划如图2所示.

2 物质科学电子实验室

要保障凝聚态物质科学数据治理的方案能够顺利实施, 数据平台与工具是必不可少的. 团队与一线科研人员密切合作, 开发了MatElab系统(图3), 并基于此系统开展了科研数据治理与应用的探索. MatElab支持实验原始数据的收集汇总, 可实现规范收集、数据安全、自定义模板填报、可检索、可定义分享、生成数据合集等功能, 可节省科研人员记录整理实验数据的时间、避免数据重复录入或缺失、便于数据统计分析等. 同时, 通过项目课题库的维护, 可将科学数据与相关项目课题进行关联, 便于数据汇交及归档管理. 软件采用自适应界面设计将WEB和移动端无缝衔接, 操作更直观、方便. MatElab可根据实际情况灵活设计, 覆盖不同专业领域的全过程研究, 包括生长制备、结构表征、物性测量全链条数据及开发优化、应用回收等相关的课题组信息及其产生的科研数据.

MatElab的主要功能模块: (1) 基本的数据管理. 数据管理分3个层级. 层级1: 记录本, 是实验记录的分类集合, 在此层级对数据进行分类、分享、合作编辑等权限管理. 层级2: 实验记录, 由若干实验数据组合生成的记录, 多条实验记录组成了一个记录本. 在此层级, 进行实验数据的检索、分类和查询. 层级3: 实验数据, 可表现为多种数据形式, 包括表单数据、表格数据、图表、文件数据、图文数据、图片集等类型. 每条实验记录由多条不同形式的数据组成; 根据数据来源可有多种生成途径, 包括日常WEB端工作中的手动录入、实验仪器通过实验记录本提供的应用程序编程接口(application programming interface, API)自动录入、外接插件生成的数据等. 实验记录本的API提供了数据的增删查改功能, 便于第三方应用的接入和用户的二次开发. (2) 预约功能. 系统提

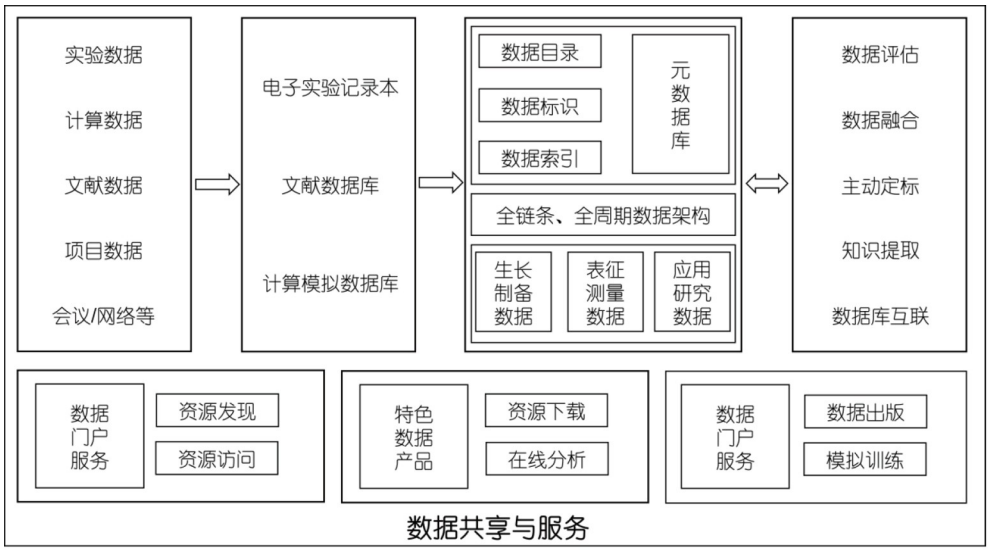


图 2 凝聚态物质科学数据治理的总体规划
Figure 2 Master plan for data governance in condensed matter science



图 3 MatElab主界面及主要功能模块
Figure 3 MatElab main interface and main functional modules

供两种预约形式: 用户自由预约; 用户提交申请设备, 管理员分配时间. 两种预约方式皆可由设备管理员自定义预约填写信息、不可预约时间段. 同时还提供设备的机时管理、预约权限管理、设备权限管理, 提升设备管理方式的自由度, 适配不同课题组的需求. (3) 模板功能. 可根据不同实验任务定制不同模板, 统一记录本格式, 减少用户重复创建模块的工作. 新的记录仅需要在模板的基础上修改少量参数, 就可生成新的数据. 团队用户可通过分享模板实现实验记录格式的统一, 减少重复工作, 提高数据可读性. (4) 数字签名. 实验记录本设计了数字签名方式对每条实验数据进行签名. 数字签名采用RSA(Rivest-Shamir-Adleman)非对称加密技术和时间戳服务双重保护数据, 提供了对数据完整性的验证功能和最后修改时间的确认. (5) 协同编辑. 支持多用户同时编辑同一记录并同步显示最新的内容. (6) 数据分布式存储. 数据的存储方式分为两种: 一种是集中式存储, 不同用户可以把数据集中存储在数据中心提供的服务器中, 由数据中心提供数据的保管、备份等管理工作; 二是分布式存储, 用户可以维护自己的数据服务器, 自己管理自己的数据. 多种存储方式可实现不同用户对数据的不同管理要求. (7) 插件管理. 针对数据处理方式不同的需求, 实验记录本开发了插件功能. 通过编写和部署不同功能的插件, 可以实现对记录本中数据的自定义处理能力. (8) 数据合集. 可将记录本中有价值、经过提炼的数据进行分类整理, 生成新的数据集合, 是在记录本的基础上进行数据整理的工具, 便于对科学数据进行统计及分析. (9) 人员管理. 实验记录本的人员和组织机构信息同步中国科学院资源规划项目(Academia Resource Planning, ARP), 可自行选择添加常用联系人, 快速完成记录本及模板的定向分享. (10) 课题库管理. 通过课题库的维护, 可逐一将科学数据与相关课题进行关联, 便于数据汇交及归档管理. (11) 论文管理. 用户上传论文文件, 自动获取论文相关信息. 同时提供了实验记录和论文的关联功能, 将论文与支撑论文的数据联动管理. (12) 数据汇交. 实验记录本提供了把实验数据汇交到中国科学院凝聚态物质科学数据中心、科学数据银行(Science Data Bank, Science DB)等数据中心的功能.

3 利用MatElab实现数据规范化

MatElab提供一个高自由度的数据操作平台. 类似一个操作系统, 安装什么软件、什么用途都可由用户自定义. 因此, 科研人员可根据其研究领域特点和需要来设计使用的规则和方法. 利用MatElab实现数据规范化, 从数据角度来看, 应考虑数据收集、数据存储、数据处理、数据应用、数据回溯、数据存档和数据销毁整个生命周期; 从科研人员的角度来看, 需考虑制备加工、测量表征、分析总结等以及相关的设备维护、论文发表、成果汇交的科研流程. 图4给出了一种利用MatElab实现科研数据规范化和科研课题联动管理的设计流程: 各课题组根据实际需求建立统一数据模板, 利用模板分享功能, 先统一课题组内数据收集和处理的格式, 组内人员在实验过程中建立数据记录, 按照模板要求填写数据. 在实际场景中, 数据收集可采用手动和自动两种方式: 手动方式是实验人员通过网页端填写记录和数据; 自动方式是通过调用API接口, 仪器自动上传实验数据. 课题组可利用插件功能, 自主开发数据处理方式, 经过统一的插件处理数据, 将结果反馈到记录本, 实现数据收集、数据处理、数据结果的全方位管理. 实验链条完成后, 定稿实验记录, 则此纪录将不可再修改, 定稿会验证数字签名并对数据加盖时间戳, 确保数据的可追溯性.

在完成数据的收集、整理、分析并发表论文后, 可进一步通过MatElab实现论文、课题、档案的管理, 用户上传论文文件, 系统自动提取论文信息, 用户可将支撑该论文的实验数据与之进行关联. 对于课题, 用户可建立课题项目, 添加课题成员, 每位成员可将与课题相关的实验数据、论文数据、文件材料进行关联, 形成完整的数据链条, 便于课题结题和汇交. 同时, MatElab可对多类型材料管理, 系统已按照中国科学院档案部门的要求, 提前预定义科研项目过程电子文件分类, 用户可将对应文件上传, 亦可自定义数据分类, 实现文件档案的高效管理, 同时系统提供文件权限管理功能.

MatElab与中国科学院凝聚态物质科学数据中心、Science DB合作, 可在MatElab一键汇交数据. 目前, MatElab开发团队在进行各种项目汇交工作的调研, 努力实现能在MatElab完成大部分类型项目和课题的数据汇交工作, 减轻科研人员项目管理的压力.

4 MatElab辅助物质研究案例

基于MatElab工具设计构建了一系列实验材料科学研究的流程后, 需要真正地应用到科研工作中来检验效果. 以下列举几个案例来展示在物质科学范畴下不同领域进行数据规范化探索所作的尝试和呈现的效果.

4.1 大幅提高非晶合金新材料的研发效率

中国科学院物理研究所柳延辉课题组[52]基于“材料基因组”思想, 利用高通量实验建立高质量、数据可比性强的非晶合金数据库, 为大数据与人工智能方法提供成分与结构的大数据. 为获取高质量数据, 他们利用MatElab将样品制备、测试、数据处理3部分融合在一起, 形成了一个规范统一、数据链条完整的实验记录. 通过利用规范实验记录形成的高质量数据库成功发现了一种非晶结构与非晶形成能力的定量关系, 并成功开发出多种非晶合金新材料[52]. 此外, 这些均衡的实验数据弥补了过去的文献通常只报道合成的非晶成分, 而忽略了晶体成分, 导致机器模型缺少负样本数据可学习的遗憾. 他们在高通量X射线衍射实验中, 表征出了同一组合金系下的非晶和晶体成分, 并利用MatElab完整记录数据, 这为机器学习提供了均衡的实验数据. 采用支持向量机模型,

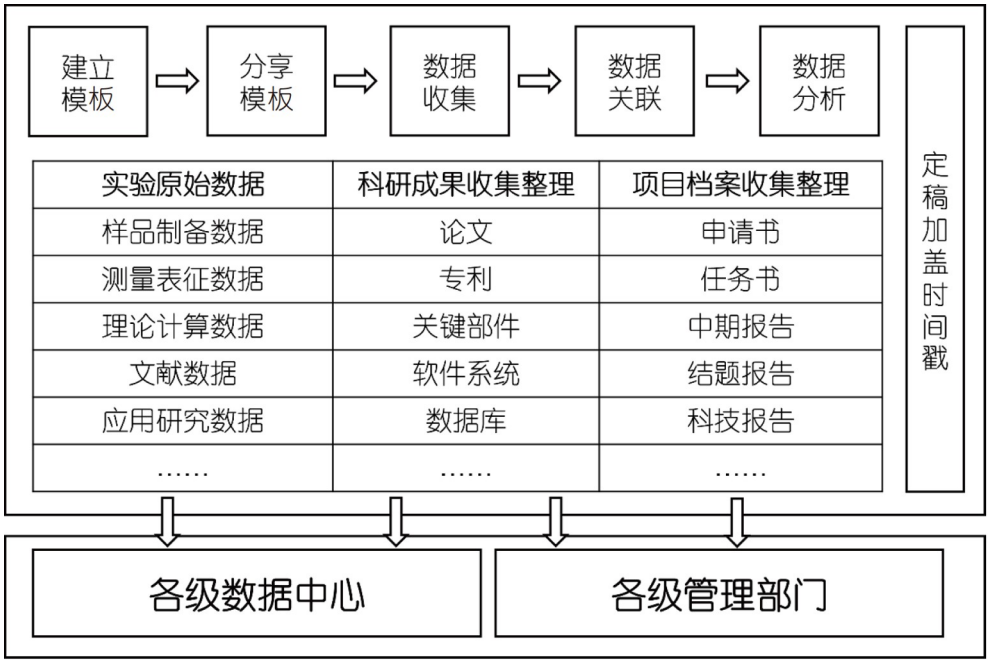


图 4 利用MatElab实现科研数据规范化和科研课题联动管理的设计流程

Figure 4 The design process of realizing standardized management of research data and linkage management of scientific research projects by means of using MatElab

优化分类参数，通过高质量数据增量的迭代，模型分类准确率持续提升. 利用该模型，成功预测并验证了多种非晶合金新材料[52]. 在这项工作中，高质量数据对模型的提升起到了至关重要的作用.

4.2 提升高质量单晶薄膜生长的良品率

自高温超导电性被发现以来，其微观机理的理解是一个长期的挑战. 在揭示高温超导机理的过程中，构建精细而系统的相图是必不可少的. 高温超导体多为多元化合物，材料制备过程中实验参数的扰动可能会在样品组分中引起弥散，进而改变其物性. 因此，研究的难点之一就是样品重复率不高，使得不同的实验批次难以得出一致的结论. 在多参量材料系统中，通过传统实验记录方式收集数据进行分析，既繁琐又耗时.

$FeSe_xTe_{1-x}$是铁基超导体中结构简单、物性丰富的典型材料. 因Se具有较强的挥发性，使得$FeSe_xTe_{1-x}$材料中实际的Se组分显著偏离了名义组分，导致不同方法甚至同一方法不同批次合成得到的样品性能存在较大的波动，如之前不同研究团队给出的该材料的相图并不完全一致[53,54]. 针对单一变量控制精度的难题，Lin等人[55]利用组合激光分子束外延技术，在一片衬底上生长出x连续变化的$FeSe_xTe_{1-x}$单晶薄膜，其中x实现了从0到1的覆盖，并结合高通量表征技术给出了组分分辨率优于0.01的$FeSe_xTe_{1-x}$输运相图. 更为重要的是，他们利用MatElab提供的数据接口，通过编程实现实时监测记录整个激光镀膜过程中包括环境湿度、气压等在内的60多项参数，信息量远超传统的手工记录. 这些详实而又全面的数据可用于对比发现不同批次实验间的差异，其数据分析与实验所观察到的现象一致[56,57]. 可见，先进实验技术与MatElab相结合，自动化地收集实验数据有助于揭示影响样品特性的关键参量，获得准确的定量关系和快速建立二维甚至高维相图. 这对于需要精确成分控制的量子材料的研究十分重要.

4.3 角分辨光电子能谱图像处理案例

为从谱图中提取出更多的物理信息，降低常见噪声处理方法过度处理而造成本征信息丢失的可能性，中国科学院物理研究所团队将机器学习与深度学习技术应用在谱图处理中，能够在尽可能保留特征的前提下提升谱图的质量. 谱图作为一个高维数据，往往会存在某种低维结构，可以利用这些低维结构实现噪声和网格的去除[58,59]. 并借助这一特性，在不使用训练数据集的情况下，利用卷积神经网络对实验数据进行处理. 目前, 该谱图处理方法已被嵌入MatElab, 用户可在线去除谱图的噪声和网格. 同时，利用电子记录本不断积累高质量的数据集后，卷积神经网络对谱图的处理效果会不断提升, 从而提取出更多细节的信息, 助力物理规律的发现.

4.4 加速电池材料改性策略的设计

锂离子电池是面向未来可充放电能源技术的关键，其正

极材料的组成及稳定性决定了电池的容量和循环性能. 锰酸锂正极作为已实现商业化的正极材料之一, 具有不可替代的作用, 但伴随着循环次数的增加而发生的材料失效为电池性能的提升带来了巨大挑战. 通过将高通量计算和典型实验设计相结合, 确定了锰溶出倾向于发生在高充电态下后, 且锰溶出是包含电解液分解和表面重构在内的表界面演化的结果, 其核心因素是表面失氧和碳酸乙烯酯(ethylene carbonate, EC)分解. 在明确了表面失氧和EC分解对于锰酸锂正极中锰溶出的关键作用后, 对尖晶石$LiMn_2O_4$表面进行Mg、Ca、Sr、Al、Ga、In、Nb、Ru、Mo、V、Tc、Ti等元素掺杂的高通量计算, 评估了各种掺杂方式对表面羟基形成和对表面失氧反应的影响, 确定了Mo、Mg、Nb对抑制Mn溶解的正面效果, 并通过实验对上述结果进行了验证. 该过程收集的数据为正极材料中氧稳定性的机器学习模型提供了有效数据, 且为二维模型描述因子的提取和训练集的有效构建提供了数据基础[60]. 由于掺杂量方案繁杂、掺杂量低、影响因素多, 利用MatElab根据实际实验流程, 制定打磨模板, 实现样品制备、测量条件、测量步骤等数据的规范记录, 提高了数据精度, 为后续在各种改性方案中进行精确的分析比较提供可能.

4.5 仪器设备集群辅助管理案例

MatElab不仅能以某一实验为对象记录实验过程中产生的数据, 还能以一个仪器设备集群为对象. MatElab提供可定义的机时预约界面, 自动收集整理仪器设备的运转情况、状态信息等. 一方面, 为用户了解仪器状态、设备机时提供了平台, 也为仪器工程师及时有效维护设备提供了信息; 另一方面, 团队负责人也可以根据从MatElab得到的设备运转效率、成果产出等统计信息来调配资源和人员, 进而提高设备使用率和科研效率.

在怀柔材料基因平台单晶薄膜实验站, 站内的8套镀膜设备和4套薄膜表征设备都可根据实际需求设置预约模式. 同时, 每台仪器设备均与MatElab通过接口相联, 实时地将设备的各项参数上传. 用户被授权后, 可以在异地登录MatElab了解设备的状态, 管理员可以根据用户的需求设定用户可见的内容. 图5为展示给实验室负责人的界面, 其中不仅包含设计的即时状态信息, 还有指定时间段产生的样品类型、数量等统计信息. 基于这套系统, 该实验站已实现了电子预约、数据提交规范化管理. 在半年内产生的数据量约为20 GB, 是传统仪器管理和数据记录方式5年总数据量的40倍.

5 总结与展望

物质研究的新范式需要规范化的高质量数据, 这需要科学家团体的共同努力, 也需要尽早地达成共识来制定数据存储和共享的标准. 我国在这方面起步较晚, 例如, 在计算工具平台方面, 目前我国尚无与世界知名材料数据工具平台(如美国Materials Project[61]、Aflow[62]和德国NOMAD[63]等)可比拟的材料科学数据平台, 这导致我国的材料、物理、化学科研都长期依赖欧美的材料科学数据生态, 由此带来诸多制约和数据安全隐患. 因而中国也迫切需要探索自己的科研数据规范化路线和发展自主的科研数据平台. 中国科学院凝聚态物质科学数据中心团队进行了大胆的探索, 从一线科研人员的实际需求出发, 针对科研数据传统流转过程中的难点和痛点,



图 5 实验站设备实时状态监测和运行统计界面
Figure 5 Real-time status monitoring and statistics interface of experimental station instruments

研发了MatElab系统. 该系统包括了数据的规范收集、安全存储、关联处理、开放共享、自动汇交等主要功能模块, 覆盖了科研数据的全生命周期. 同时, 主要的功能模块还提供了可供二次开发的接口, 满足个性化功能的需求. 目前, MatElab已在中国科学院物理研究所的部分课题组投入使用, 并推广至怀柔科学实验平台和大装置. 科研人员将这些工具和思路在不同的科研场景中加以应用, 在大幅提高研发非晶合金新材料的效率, 及快速提升功能材料单晶薄膜质量、角分辨光电子能谱谱图智能去噪和材料数据库建设等工作中发挥了重要的作用. 一系列成功的案例可供后继工作参考. 为在物质科学领域数据的规范化探索出一条具备可操作性的路线.

需要指出的是, 科研数据治理只是大数据科学发现研究范式的基础, 在后期仍需继续开展元数据、数据目录、数据标识、数据开放共享等工作, 其目的是希望数据能够更加有效地用于后继科研, 如在复杂的参量体系中搜寻物理规律等. 但规范化后, 数据的用途不仅限于此, 如, (1) 将实验记录本预约功能与记录功能相结合, 进一步完善软硬件接口, 建立仪器设备与MatElab的双向数据通信, 可实现大型实验室的自动化管理; (2) 利用MatElab的合作编辑功能, 将课题攻关任务模块化、节点化, 项目内成员可以实时分享最新产生的数据或进展, 项目负责人可以第一时间了解项目/课题的进展情况, 以便及时合理地调配资源; (3) 相关数据进一步与人工智能结合, 实现实验室的智能化管理, 将科研人员从繁复的机械式劳动中解脱出来, 可以将更多的精力投入思考或设计等创造性工作; (4) 以上是在实验室的层面, 如果扩大到某一个科研领域, 高质量数据的分享可以减少科研资源的浪费, 提高整个领域的产出投入比. 相关的模块已在MatElab设计及后继试用中初现端倪, 但完整功能的实现还需要开展大量的工作: 如打造智慧实验室, 需要设计大量不同类型仪器设备与数据终端通信的硬件接口和协议标准, 建立包含纠错功能的多级安全仪器控制机制等; 营造科研数据分享的生态环境需要提供大量针对科研数据库的应用程序, 以满足不同场景下的需求, 并建立数据安全分享和交易的机制, 建立科研工作者数据信用系统等. 全生命周期的科研数据管理和利用任重而道远, MatElab只走出了一步, 未来仍需各方面力量共同努力, 持续发展下去, 才能充分彰显基于大数据新研究范式的变革力量.

致谢 感谢国家重点研发计划(2022YFA1603903, 2022YFA1403800, 2021YFA0718700, 2022YFA1403900)、国家自然科学基金(11927808, 12225412, 52022106, 11925408, 11921004, 12188101, 12374141, 12274439)、国家自然科学基金优秀青年科学基金(T2222028)、中国科学院B类战略性先导科技专项(XDB33000000, XDB25000000)和中国科学院网络安全和信息化专项(CAS WX2021SF-0102)资助.

推荐阅读文献

1 Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data. Nature, 2009, 457: 1012–1014
2 Abdella J A, Zaki N M, Shuaib K, et al. Airline ticket price and demand prediction: A survey. J King Saud Univ Sci, 2021, 33: 375–391
3 Van Dijck J. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. Surveill Soc, 2014, 12: 197–208
4 Prakash A, Sandfeld S. Chances and challenges in fusing data science with materials science. Prakt Metallogr-Pract Metallogr, 2018, 55: 493–514
5 Hu X F, He X Y, Xu X L. Challenges and reflections on modeling and simulation in the era of Big Data (in Chinese). Sci Sin Inform, 2014, 44: 676–692 [胡晓峰, 贺筱媛, 徐旭林. 大数据时代对建模仿真的挑战与思考. 中国科学: 信息科学, 2014, 44: 676–692]
6 Shen Y, Brandt W N, Dawson K S, et al. The sloan digital sky survey reverberation mapping project: Technical overview. Astrophys J Suppl Ser, 2015, 216: 4–29
7 Zhan H, Tyson J A. Cosmology with the Large Synoptic Survey Telescope: An overview. Rep Prog Phys, 2018, 81: 066901
8 Bansal M. DNA structure: Revisiting the Watson-Crick double helix. Curr Sci, 2003, 85: 1556–1563
9 Liu L, Li Y, Li S, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol, 2012, 2012: 251364
10 Stein L D. The case for cloud computing in genome informatics. Genome Biol, 2010, 11: 207
11 Qin X P, Wang H J, Du X Y, et al. Big Data analysis—Competition and symbiosis of RDBMS and MapReduce (in Chinese). J Software, 2012, 23: 32–45 [覃雄派, 王会举, 杜小勇, 等. 大数据分析——RDBMS与MapReduce的竞争与共生. 软件学报, 2012, 23: 32–45]
12 Arbey A, Battaglia M, Djouadi A, et al. The Higgs sector of the phenomenological MSSM in the light of the Higgs boson discovery. J High Energy Phys, 2012, 2012: 107
13 Thomas S K. Jin W L, Hu X H, trans. The Structure of Scientific Revolution (in Chinese).Beijing: Peking University Press, 2012 [托马斯·塞缪尔·库恩. 金吾伦、胡新和, 译. 科学革命的结构. 北京: 北京大学出版社, 2012]
14 Sha W, Guo Y, Yuan Q, et al. Artificial intelligence to power the future of materials science and engineering. Adv Intell Syst, 2020, 2: 1900143
15 Shen L, Bai J, Wang J, et al. The fourth scientific discovery paradigm for precision medicine and healthcare: Challenges ahead. Precis Clin Med,

2021, 4: 80–84
16 Tolle K M, Tansley D S W, Hey A J G. The Fourth Paradigm: Data-intensive scientific discovery. Proc IEEE, 2011, 99: 1334–1337
17 Rodrigues Jr J F, Florea L, de Oliveira M C F, et al. Big Data and machine learning for materials science. Discov Mater, 2021, 1: 12
18 Gil Y, Greaves M, Hendler J, et al. Amplify scientific discovery with artificial intelligence. Science, 2014, 346: 171–172
19 Pinsky E. Teaching data science by history: Kepler's laws of planetary motion and generalized linear models. In: Computer Science and Education Conference, 2020. 42443
20 Laughlin R B, Pines D. The theory of everything. Proc Natl Acad Sci USA, 2000, 97: 28–31
21 Brunin G, Ricci F, Ha V A, et al. Transparent conducting materials discovery using high-throughput computing. npj Comput Mater, 2019, 5: 63
22 Wang H Z, Wang H, Ding H, et al. Progress in high-throughput materials synthesis and characterization (in Chinese). Sci Technol Rev, 2015, 33: 31–49 [王海舟, 汪洪, 丁洪, 等. 材料的高通量制备与表征技术. 科技导报, 2015, 33: 31–49]
23 Gomes C P, Selman B, Gregoire J M. Artificial intelligence for materials discovery. MRS Bull, 2019, 44: 538–544
24 Xu G, Jiang R, Wei W, et al. Correlation analysis and causal analysis in the era of Big Data. IOP Conf Ser Mater Sci Eng, 2019, 563: 042032
25 Wang G J, Li K Q, Peng L Y, et al. High-throughput automatic integrated material calculations and data management intelligent platform and the application in novel alloys (in Chinese). Acta Metall Sin, 2022, 58: 75–88 [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. 金属学报, 2022, 58: 75–88]
26 Zhang T Y. An Introduction to Materials Informatics (I): The Elements of Machine Learning (in Chinese). Beijing: Science Press, 2022 [张统一. 材料信息学导论(上): 机器学习基础. 北京: 科学出版社, 2022]
27 Perdew J P. Artificial intelligence "sees" split electrons. Science, 2021, 374: 1322–1323
28 Kusne A G, Yu H, Wu C, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun, 2020, 11: 5966
29 Fung V, Zhang J, Hu G, et al. Inverse design of two-dimensional materials with invertible neural networks. npj Comput Mater, 2021, 7: 200–209
30 Zhu S Q, Yu T, Xu T, et al. Intelligent computing: The latest advances, challenges, and future. Intell Comput, 2023, 2: 0006
31 Ozdemir V, Smith C, Bongiovanni K, et al. Policy and data-intensive scientific discovery in the beginning of the 21st century. OMICS-J Integr Biol, 2011, 15: 221–225
32 Guo J, Chen Z, Liu Z, et al. Neural network training method for materials science based on multi-source databases. Sci Rep, 2022, 12: 15326
33 Kalidindi S R, De Graef M. Materials data science: Current status and future outlook. Annu Rev Mater Res, 2015, 45: 171–193
34 Zhang Q, Wu D Y, Zhao J H. Big Data standards system (in Chinese). Big Data Res, 2017, 3: 11–19 [张群, 吴东亚, 赵菁华. 大数据标准体系. 大数据, 2017, 3: 11–19]
35 DAMA. Data Management Body of Knowledge (in Chinese). Beijing: China Machine Press, 2022 [DAMA. 数据管理知识体系指南. 北京: 机械出版社, 2022]
36 Banko L, Ludwig A. Fast-track to research data management in experimental material science—Setting the ground for research group level materials digitalization. ACS Comb Sci, 2020, 22: 401–409
37 Soedarmadji E, Stein H S, Suram S K, et al. Tracking materials science data lineage to manage millions of materials experiments and analyses. npj Comput Mater, 2019, 5: 79
38 Zakutayev A, Wunder N, Schwarting M, et al. An open experimental database for exploring inorganic materials. Sci Data, 2018, 5: 180053
39 Kube S A, Sohn S, Uhl D, et al. Phase selection motifs in high entropy alloys revealed through combinatorial methods: Large atomic size difference favors BCC over FCC. Acta Mater, 2019, 166: 677–686
40 Park W B, Chung J, Jung J, et al. Classification of crystal structure using a convolutional neural network. IUCrJ, 2017, 4: 486–494
41 Zaloga A N, Stanovov V V, Bezrukova O E, et al. Crystal symmetry classification from powder X-ray diffraction patterns using a convolutional neural network. Mater Today Commun, 2020, 25: 101662
42 Choudhary K, DeCost B, Chen C, et al. Recent advances and applications of deep learning methods in materials science. npj Comput Mater, 2022, 8: 59
43 Higgins S G, Nogiwa-Valdez A A, Stevens M M. Considerations for implementing electronic laboratory notebooks in an academic research environment. Nat Protoc, 2022, 17: 179–189
44 Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data, 2016, 3: 160018
45 Kwok R. How to pick an electronic laboratory notebook. Nature, 2018, 560: 269–270
46 Kanza S, Willoughby C, Gibbins N, et al. Electronic lab notebooks: Can they replace paper? J Cheminformatics, 2017, 9: 31
47 Hewera M, Hänggi D, Gerlach B, et al. eLabFTW as an open science tool to improve the quality and translation of preclinical research. F1000Res, 2021, 10: 292
48 Xiong S L, Wang L H. Research progress and development trends of materials genome technology. Adv Mater Sci Eng, 2020, 2020: 5903457
49 Xiang X D, Wang H, Xiang Y, et al. Applications of combinatorial material chip technology in research and development of new materials (in

Chinese). Sci Technol Rev, 2015, 33: 64–78 [项晓东, 汪洪, 向勇, 等. 组合材料芯片技术在新材料研发中的应用. 科技导报, 2015, 33: 64–78]
50 Pasquetto I V, Randles B M, Borgman C L. On the reuse of scientific data. Data Sci J, 2017, 16: 1–9
51 Wang D, Zhou M B, Li Y L. Research on automated data acquisition technology of thin film growth. Front Data Comput, 2023, 5: 29–38
52 Li M X, Sun Y T, Wang C, et al. Data-driven discovery of a universal indicator for metallic glass forming ability. Nat Mater, 2022, 21: 165–172
53 Zhuang J, Yeoh W K, Cui X, et al. Unabridged phase diagram for single-phased $FeSe_xTe_{1-x}$ thin films. Sci Rep, 2014, 4: 7273
54 Imai Y, Sawada Y, Nabeshima F, et al. Suppression of phase separation and giant enhancement of superconducting transition temperature in $FeSe_{1-x}Te_x$ thin films. Proc Natl Acad Sci USA, 2015, 112: 1937–1940
55 Lin Z, Tu S, Xu J, et al. Phase diagrams on composition-spread $Fe_ySe_{1-x}Te_x$ films. Sci Bull, 2022, 67: 1443–1449
56 Zhang Y, Wang T, Wang Z, et al. Effects of Te- and Fe-doping on the superconducting properties in $Fe_ySe_{1-x}Te_x$ thin films. Sci Rep, 2022, 12: 391
57 Feng Z, Yuan J, He G, et al. Tunable critical temperature for superconductivity in FeSe thin films by pulsed laser deposition. Sci Rep, 2018, 8: 4039
58 Liu J, Huang D, Yang Y, et al. Removing grid structure in angle-resolved photoemission spectra via deep learning method. Phys Rev B, 2023, 107: 165106
59 Huang D, Liu J, Qian T, et al. Spectroscopic data de-noising via training-set-free deep learning method. Sci China-Phys Mech Astron, 2023, 66: 267011
60 Sun X, Xiao R, Yu X, et al. Screening $LiMn_2O_4$ surface modification schemes under theoretical guidance. ACS Appl Mater Interfaces, 2022, 14: 10353–10362
61 Jain A, Ong S P, Hautier G, et al. The materials project: A materials genome approach to accelerating materials innovation. APL Mater, 2013, 1: 011002
62 Calderon C E, Plata J J, Toher C, et al. The AFLOW standard for high-throughput materials science calculations. Comput Mater Sci, 2015, 108: 233–238
63 Wang C C, Su H, Duan L N, et al. Material data sharing and potential application of blockchain: A review and discussion (in Chinese). Mater Rep, 2022, 36: 21020155 [王畅畅, 苏航, 段琳娜, 等. 材料数据共享现状综述及区块链应用前景探讨. 材料导报, 2022, 36: 21020155]

Summary for “凝聚态物质科学科研数据管理与应用”

Management and application of research data in condensed matter science

Dan Wang[1,2,4†], Mingbo Zhou[1,2†], Dongchen Huang[1,2], Yunlong Li[3], Zefeng Lin[1], Junde Liu[1], Tiannian Zhu[1,2], Yun Zhu[4], Mingxing Li[1,2], Ruijuan Xiao[1,2], Jie Yuan[1,2*] & Hongming Weng[1,2*]

[1] *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China;*
[2] *Condensed Matter Physics Data Center, Chinese Academy of Sciences, Beijing 100190, China;*
[3] *Bureau of Frontier Science and Education, Chinese Academy of Sciences, Beijing 100864, China;*
[4] *College of Physics and Materials Science, Tianjin Normal University, Tianjin 300387, China*
† Equally contributed to this work
* Corresponding authors, E-mail: yuanjie@iphy.ac.cn; hmweng@iphy.ac.cn

Massive amounts of data, dramatically growing computing power, and the development of the digital economy continue to give rise to data science. The rapid development in technologies, such as machine learning, artificial intelligence, and blockchain, has enhanced this trend. Data have become a new factor of production, bringing revolutionary changes to people’s lives and production techniques. In this context, scientific research has also begun to shift towards a new data-driven paradigm, using massive amounts of data as a basis to reveal the correlations hidden behind them, thereby adding new dimensions and perspectives to existing conventional research.

Recently, data-based technologies, such as artificial intelligence, which are based on Big Data, have been gradually involved in scientific research. These technologies are used for integrating theory, computation, and experimental measurements, injecting new direction and impetus into scientific research. For example, algorithms such as random forests and neural networks have become commonly used data processing and analysis tools in materials science and achieved remarkable results in predicting properties, phase diagrams, and structures. Traditional scientific research methods often rely on specific theoretical assumptions and experimental designs, while data-driven scientific research focuses on obtaining knowledge and insights from data. By mining hidden correlations and patterns from massive amounts of data, researchers can conduct exploratory studies, discover new research directions and questions, expand the boundaries of research fields, and bring novel insights and breakthroughs to their research.

Due to the importance of data, higher requirements for data management and utilization are demanded in scientific research. Currently, scientific research data is hindered by diverse formats, fragmented distribution, and inconsistent quality, causing data to remain inside the laboratory; this results in the wastage of scientific research resources and declination in the development of scientific research. To effectively manage data, improve data utilization, and address the issue of “data islands”, researchers have developed an Electronic Laboratory for Material Science, a data management platform based on the actual needs of frontline researchers and difficulties associated with the traditional flow of scientific research data. This platform covers the entire lifecycle of data production, collection, storage, analysis, and sharing. Furthermore, it is committed to realizing planned information processing, automated information collection, intelligent data empowerment, and promoting the commencement of a new paradigm in scientific research.

This article introduces the development of scientific research paradigms and current advancements in data management at home and abroad, emphasizing the importance of data standards in scientific research and highlighting the successful exploration of the Electronic Laboratory for Material Science for significantly improving the efficiency in developing amorphous alloy materials, enhancing the quality of functional material single-crystal films, intelligently denoising angle-resolved photoemission spectroscopy spectra, and assisting the management of large-scale experimental stations. With this data management platform, researchers can now more effectively organize and manage experimental data, ensuring its integrity, systematicity, and standardization, thereby improving the efficiency and accuracy of their scientific research work and providing feasible paths and exemplary cases for data management in materials science.

Based on the successful application of the Electronic Laboratory for Material Science, the research and development team will further improve its function and build an independent data platform for scientific research in China. In the era of Big Data, open sharing data and establishing an information-based ecosystem will establish the foundation for developing a smart laboratory and injecting new vitality into materials science advancements.

research paradigm, intensive data mining, Electronic Laboratory for Material Science, data normalization

doi: 10.1360/TB-2023-0463