引用格式：陈凯华, 刘泓欣, 郭锐. 我国加强支持科研智能数据库建设的意义、挑战与建议. 中国科学院院刊, 2026, 41(1): 194-204, doi: 10.3724/j.issn.1000-3045.20250423004.
Chen K H, Liu H X, Guo R. Significance, challenges, and policy recommendations for strengthening database development to support AI for Science in China. Bulletin of Chinese Academy of Sciences, 2026, 41(1): 194-204, doi: 10.3724/j.issn.1000-3045.20250423004. (in Chinese)

我国加强支持科研智能数据库建设的意义、挑战与建议

陈凯华[1] 刘泓欣[1] 郭 锐[2*]

1 中国科学院大学 公共政策与管理学院 北京 100049
2 北京理工大学 经济学院 北京 100081

摘要 以大数据和人工智能为驱动的科研智能迅速兴起，高质量、开放共享的科学数据库建设愈加成为科技创新、提升科技竞争力的战略制高点。世界主要国家愈发重视科学数据库建设在基础研究发展中的支柱性作用，在不断加强高质量科学数据库建设的同时，持续对我国包括科研数据在内的一众人工智能技术及所需资源的发展进行封锁打压。在此背景下，如何构建起我国自主可控的支持科研智能发展的科学数据生态体系，对于发掘我国科学数据潜力、保障国家科研安全、实现自主创新能力跨越式提升具有重要意义。在该背景下，文章回顾了国际上支撑科研智能相关数据库建设的经验，分析了我国科研智能数据库建设现状与问题，识别出包括科学数据存储数量与质量不匹配、开源开放与共享不足、政策支持体系不健全且协同不足，以及数据库应用生态尚不成熟等多方面挑战，在此基础上从加强顶层设计和政府支持、推动开源开放和标准化管理、促进多元合作和与产业场景应用对接等方面提出加强我国科学数据生态体系建设的建议。

关键词 科研智能，科学数据库，自主可控，开源开放，标准化管理

DOI 10.3724/j.issn.1000-3045.20250423004
CSTR 32128.14.CASbulletin.20250423004

科研智能（AI for Science，AI4S）在支持重大科学发现和颠覆性创新中的基础性和工具化属性日益突出，国际科研智能竞争日益白热化。高质量的科学数据作为影响科研智能发展质量和水平的关键要素，已

*通信作者

资助项目：国家杰出青年科学基金项目（72025403），国家自然科学基金面上项目（72574020），中国科学院战略研究与决策支持系统建设专项（GHJ-ZLZX-2025-11），国家自然科学基金专项项目（L25240113），中国科协青年人才托举工程（2023QNRC001）
修改稿收到日期：2026年1月6日

成为全球科技竞争的战略制高点。区别于传统数据集的构建，科研智能发展对于相关科学数据库构建提出了更高的要求，需要更高质量、跨学科、大存量、多模态的科学数据[1]，以满足人工智能模型训练的多样化和高性能需求；同时，需要相关科学数据存储具备全链条、完整性及动态更新的特征，并具有强大的算力、算法支持，具备良好的互操作性和兼容性，能够支撑数据共享和交互，同时为数据库使用者提供强有力的科研智能数据获取和处理服务支持[2]。

目前，我国科学数据库建设总体仍处于起步阶段，与高水平科研所需求的科学数据库相比仍存在差距。与此同时，以美国为代表的西方国家在科技领域持续推进单边主义做法，对包括人工智能在内的多项前沿关键技术实施限制，并不断收紧对我国科研数据的获取与使用条件，相关外部约束呈现加剧态势。2024年2月28日，美国白宫发布了14117号行政命令，限制受关注国家的数据获取权；在此基础上，美国司法部于2024年12月27日发布了《防止受关注国家及相关人员访问美国敏感个人数据和政府相关数据的规定》（*Preventing Access to U.S. Sensitive Personal Data and Government- Related Data by Countries of Concern or Covered Persons*）。美国国立卫生研究院（NIH）也宣布自2025年4月4日起禁止位于中国、俄罗斯、伊朗等受关注国家的机构访问NIH部分数据存储库和相关数据，涉及多个生物领域重要的科学数据库和平台，这对我国生物医药的发展产生了较大的冲击，也将极大掣肘我国科研智能的发展。

相较于美欧国家和地区对高质量科学数据库体系建设提供的强有力支撑，我国支持科研智能发展的数据库建设仍面临存储数量与质量不匹配、开源开放与共享不充分、政策支持体系协同不健全、数据库应用生态尚不成熟等挑战。因此，亟须借鉴国际相关数据体系建设的成功经验，从加强顶层设计和政府支持、推动开源开放和标准化管理、促进多元合作和实现与产业场景应用对接等方面提供针对性的强化措施，以支撑我国科研智能发展及国际竞争能力的提升。

1 构建支撑科研智能发展的科学数据库的重要意义

推动支持科研智能的科学数据库体系构建对于挖掘我国科学数据资源潜力、打造可信数据空间、形成发展合力以支撑我国情境下科研智能发展具有重要意义。

1.1 挖掘我国科学数据资源潜力，赋能国家科研智能发展

当前，我国科学数据虽然存量巨大，但存在开发低效、资源浪费严重、管理存储不当、整合壁垒限制等问题，致使大量科学数据闲置或流失，难以有效转化为可用资源。科研智能数据库建设过程也是促进研究过程中数据记录标准化、数据管理专业化和体系化的过程。通过进行科研智能数据库建设，可以充分挖掘和收集高质量、大存量的科学数据资源，并且进一步健全我国科学数据管理制度，整合利用我国现有科学数据资源，进一步发挥科学数据资源潜力。此外，相较于国外数据资源，以本国情境下的科研智能数据为支撑，能够为我国科研智能发展提供所需要的数据资源支持，赋能本国情境下科研智能的发展。

1.2 打造科研智能发展的可信数据空间，形成发展合力

科研智能发展及其相关数据库建设强调跨学科、跨领域、跨国家合作的开展，但现阶段我国科研智能发展及科研智能数据库建设存在着主体间合作壁垒高、信任机制不健全，以及科学数据权益保护机制不足等问题，这些问题限制了科研智能发展合力的形成。而构建科研智能数据库体系需要建立一个安全、可信的数据共享和流通环境[1]，以支持科研数据的高效利用和智能化应用，同时确保数据的可信性、隐私性和合规性，以提升各主体的科学数据共享意愿。在

过程中同步加强合作，进行合作机制设计、合作交流平台构建及可信隐私保护技术支持等，以促进不同领域学科数据汇集，并进一步促进各科研智能研究者合作，形成科研智能发展合力。

1.3 掌握科研智能发展自主权，提高国际科技竞争力

目前，我国科研智能研究很大程度上依赖于国外的大模型及科学数据库等，致使我国科研智能发展的自主性受限。一旦国外政府对我国研究人员的访问和应用权限进行限制，将会极大掣肘我国科研智能及科学技术创新的发展。在当前美国强力限制我国科研人员对其科学数据库访问权利的背景下，我国科研智能发展的形势愈加严峻。例如，据我国国家生物信息中心估算，2025年4月美国NIH对我国科研人员的数据访问限制可能导致我国创新药企的研发效率下降20%—30%，进而影响我国创新药企的国际市场竞争力。因此，构建我国自主的科研智能发展数据库，是掌握科研智能发展自主权，破除国际垄断，提升我国科研智能发展国际竞争力的重要举措。

2 支撑科研智能相关数据库建设的国际经验

AlphaFold等科研智能模型之所以能够最早出现于美国，离不开其成熟、健全的科学数据库体系的支持。例如，AlphaFold的重要训练数据来源之一蛋白质数据库（PDB）于1971年由美国布鲁克海文国家实验室建立，是全球首个生物大分子结构数据库，也是生物科学中第一个开放获取的数据资源。美国国家生物技术信息中心（NCBI）自1988年起便开始进行相关科学数据库建设，经过长期积累和探索，逐步建立起了全面、完整的生物数据资源体系，并持续服务于全球科学研究与产业应用，在生命科学领域的理论探索和应用实践中发挥着不可或缺的作用。通过系统分析包括美国等在内的国家科学数据库发展情况，本文总结出包括开源开放、政府支撑、专业化标准化管理、多主体合作及动态更新等方面的建设经验。

2.1 以开源开放为核心原则

（1）加强科学数据库开放，推动科学数据库来源多样化和动态更新。通过开放访问，科研数据能够得到广泛的验证与应用，通过多方主体参与使用并进行反馈，有利于发现所存储科学数据存在的问题，在兼顾较低成本和较高效率的基础上实现科学数据质量的提升；同时能够吸引更多科研人员参与数据收集、分析和改进，促进跨学科、跨地域的合作与知识传播。例如，AlphaFold通过开源其蛋白质结构预测模型与数据，推动全球研究者的合作与数据验证；美国“材料计划”（The Materials Project）作为开放的材料科学数据库，依托全球科研人员的共同贡献，推动了计算与实验数据的共享。

（2）基于开源模式，加速科研智能数据库构建与使用生态的形成。通过对科学数据库的数据、计算平台与工具、软件等进行开源开放，同时进行多平台兼容。在此基础上鼓励支持开放社区构建，加强科研智能数据库的良性互动与应用生态建设，显著拓展数据库影响力。例如，The Materials Project对其数据库进行开源并持续进行更新，同时提供了开源的Python材料基因组学库（pymatgen）、开放材料数据库（Materials Project Contributions）等项目和工具，支持研究人员的数据处理和分析。同时构建起相应社区，并鼓励用户的参与和反馈，通过开源加速科研智能数据库构建与使用生态的形成，共同推动科学数据库的发展。

2.2 强化政府的支撑作用

（1）以制度和政策多方联动推动科研智能数据库的建设。① 以政策制度推动科学数据库开源开放[3]。美欧国家和地区高度重视科学数据的开放，发布了多项政策和指南，指引各科研主体的科学数据开放行为，形成了良好的科学数据开放的制度环境和生态。例如，美国的《确保免费、及时和公平地获得联邦资助的研究成果》（*Ensuring Free, Immediate, and*

Equitable Access to Federally Funded Research）和欧盟的“地平线欧洲计划”（Horizon Europe）等。同时，各科学资助机构也通过相关科学数据开放活动和政策，推动受资助项目进行科学数据开放。例如，美国国家科学基金会（NSF）、NIH和美国国家航空航天局（NASA）等都发布了相应的资助项目的数据公开计划；美国能源部（DOE）也要求其资助的研究项目数据必须公开，并存储于开放平台，包括能源与科学公共访问门户（DOE PAGES）①和政府开放数据（Data.gov）等，明确存储位置、周期，确保科研数据的开放与共享。NIH鼓励个人、团队、科研机构通过数据平台上传数据，并给予数据提交者一定的补贴[4]。② 以政策制度促进科学数据标准化。通过制定相关政策制度，有利于推动不同学科和行业下科学数据标准的统一，推动科学数据库建设和发展。例如，欧盟发布了《让FAIR变为现实》（*Turning FAIR into Reality*）②、《“地平线2020”项目中数据管理的FAIR指南》（*Guidelines on FAIR Data Management in Horizon 2020*）等报告与文件，要求所有参加开放研究数据先导计划的项目都必须提交符合FAIR原则的数据管理计划。③ 建立健全科学数据产权保护制度。通过明确科学数据的所属权，完善对科学数据所属主体的权益保护，能进一步激励科研主体进行科学数据共享的积极性。例如，欧盟的《通用数据保护条例》（GDPR）包括对某些数据控制者义务的具体减损，以及要求采取适当保护措施的具体规定。

（2）提供资源和基础设施保障。科研智能的发展对科学数据的数量、质量和算力等都提出了更高的要求，而单个机构或部门由于资源等的限制，难以单独支撑科研智能数据库所需的算力、存储等的需求。在该背景下，美欧国家和地区加强科研智能发展所需的资源和基础设施建设，强化对相关科学数据库建设的支撑作用。例如，美国的“国家人工智能研究资源计划”（NAIRR）、欧盟的“欧洲开放科学云”（EOSC）和“法国国家云工业计划”（Industrial plan to support the French Cloud sector）等都通过政府推动相关基础设施建设，为科研人员提供强有力的数据存储与计算资源支持，支撑了科研智能相关数据库的建设和发展。

2.3 强化科学数据库的专业化、标准化管理

（1）通过专门机构专业化做好科学数据库管理工作。国外众多成熟的科学数据库及相关机构都设立了专门的机构进行科学数据库的管理工作，通过专业人员、专业管理制度和方法，以及专业数据平台实现科学数据库的标准化、质量审查和长期维护，确保科学数据库的长期稳定运行和高效管理。例如，美国NIH资助并管理多个生物医学数据库，其中包括PubMed和GenBank等，确保了数据库的科学性与数据管理的规范性；科学技术信息办公室（OSTI）作为美国能源部下属的专门机构，负责能源科研数据的集成与管理，将来自不同领域、不同实验室的数据进行整合，为用户提供一站式访问平台，并通过平台OSTI.GOV为全球科研人员提供高效的数据查询与访问服务，从而推动了科学数据库的集成发展。

（2）强化科学数据库的标准化管理。① 标准化质量审查。通过制定统一的数据提交标准和审核流程，确保科学数据库建设过程中数据的一致性、准确性与长期可用性，以为科研智能发展提供更高质量的科学数据支持。例如，PDB采用严格的数据提交和标准化流程，经过严格的审核流程，确保上传的蛋白质结构

① DOE PAGES的全称是Department of Energy Public Access Gateway for Energy and Science（能源与科学公共访问门户），是美国能源部（DOE）的官方公共访问存储库，用于检索由能源部资助的研究产生的期刊文章和已接受稿件。

② FAIR原则是一套用于科研数据管理的国际公认指南，强调数据应当可查找（findable）、可访问（accessible）、可互操作（interoperable）和可重复使用（reusable）。其目标是提升数据的可用性和再利用价值，促进科研透明度和跨机构、跨学科的数据共享。

数据符合国际标准。同时其定期对数据格式进行更新和修复，并引入了严格的数据验证程序，以保证上传数据的高质量和一致性。② 标准化格式管理。标准化的格式管理能够有效保障数据质量，提升数据质量审查效率，促进跨学科科学数据的整合和应用，同时也有利于扩展科学数据库应用的范围，促进开源开放和生态构建[5]。例如，PubMed和PubTator等数据库依托美国NCBI统一的技术框架和标准，确保全球科研人员能够高效、准确地使用数据。

2.4 加强多主体合作推动科学数据库共建

（1）推动数据库建设过程中的跨学科跨行业合作。通过多主体合作降低学科间及行业间的科学数据融合壁垒，实现数据来源的多学科性，进一步促进交叉学科的应用与发展，同时进一步整合资源推动科学数据库共建和发展。例如，The Materials Project在建立过程中整合了不同政府机构和企业界的力量。而结合DeepMind开发的深度学习工具“材料探索图网络”（GNoME）及The Materials Project数据库预测并合成的新材料数据，进一步扩展并更新了The Materials Project。而“表观组学线图计划”（Roadmap Epigenomics Mapping Consortium）由美国NIH从2008年启动，多个研究机构和大学联合参与，通过生成表观遗传学数据集，为研究人员提供理解基因调控的数据资源。

（2）以跨国合作推动科学数据库建设的全球化。开展跨国合作，特别是在数据共享与标准化方面的国际协调，有利于构建科研智能数据库的全球生态，从而扩展数据库的使用范围，提升数据质量与一致性。例如，蛋白质数据库全球联盟（Worldwide Protein Data Bank）作为国际性合作项目，由美国、欧洲和日本的多个科研机构共同管理，确保了蛋白质数据的全球共享与标准化；同样，PDB通过这种跨国合作模式，推动了蛋白质结构数据的标准化管理和全球科研的协调。

2.5 动态更新数据资源和技术服务

在科研智能发展背景下，快速发展的人工智能技术要求相应的科学数据库能够动态更新相关数据处理技术、工具和服务，以满足进行科研智能研究的需求。同时，针对科研智能所产生的新的数据，相关科学数据库也需要建立健全相关数据审核、吸纳等机制体系，以促进自身与时俱进。国际众多成熟的科学数据库都保持着对新的技术和数据较高的敏感性和动态更新性，以及时响应科研智能发展需求。例如，开放量子材料数据库（OQMD）引入了机器学习算法，以及与多个国际研究机构合作共同开发新的材料预测模型，提高了材料性质预测的准确性和效率，进一步推动了材料科学的发展。而The Materials Project也引入了机器学习算法，并更新了相应的材料数据分析工具，旨在通过大数据和人工智能技术提升材料性质预测的准确性和效率，加速新材料的发现和开发。

3 我国支撑科研智能发展数据库的建设现状与挑战

当前，我国支持科研智能发展的数据库建设仍面临数据规模与质量不匹配、开源共享不足、政策协同机制不健全，以及应用生态尚未成熟等问题，这在一定程度上制约了我国科研智能自主发展、持续创新及其国际影响力的提升。

3.1 科学数据存储数量与质量不足，致使开发潜力有限

尽管我国科学数据积累存量较大，但由于未形成较好的数据存储和管理机制，导致所存储数据的完整性、规范性、准确性存在较大的缺失；同时，科研人员不规范的科研数据上传行为也加大了对科学数据质量审查和监管的难度，致使科学数据利用率低、整合开发不足。例如，相关调查显示，我国科研人员存在对于元数据认知不足的问题，并且在管理、利用自有和他人数据时存在诸如“文件命名混乱”“数据内容

信息不清”等管理存储不规范、数据信息缺失等问题，严重影响相关数据的共享与重用[6]。基于数据中心互联（DCI）的分析数据也显示，我国科学数据集的数量虽位居世界第2名，但是在拥有国际影响力的国际主要科学数据仓储（SDR）数量的排名中，我国排在前20名之外，科学数据的存储数量与存储质量尚不匹配[7]。

3.2 科学数据开源开放与共享不足，阻滞了资源整合与生态体系构建

（1）缺少统一管理协调机构，导致科学数据使用中的“数据孤岛”现象存在，各部门和机构各自为政，难以实现有效的科学数据整合和共享以支撑科研智能背景下高水平科学数据库的建设。例如，截至2025年9月，我国虽已建成以20个国家科学数据中心和31个国家资源库为重要基础的科技资源共享平台，但各中心分属不同的单位和部门，尚未有明确的统一管理机构对其进行协调管理，致使各科学数据中心通过网络为用户提供数据资源与服务的质量和水平参差不齐[8]。而且由于缺乏统一协调，不同机构可能重复采集相同或类似的科学数据，造成支撑科研智能发展的科学数据库在建设过程中存在资源浪费。

（2）数据标准不统一不健全，阻碍了支撑科研智能发展的高水平科学数据库。我国科研档案管理规划未包含科学数据归档格式管理内容、科学数据归档和长期保存格式标准规范有待完善[9]，且同一学科下不同机构和数据存储中心的科学数据标准要求存在异质性现象，难以进行统一化整合。而且，我国开放科学数据平台的标准治理存在数据产权与隐私模糊、关键性协同机制缺失等问题，难以支撑科学数据互联互通的实现[10]。

（3）相关科学数据开放共享的机制不健全，未能充分激发和引导科研人员进行科学数据资源共享的积极性。例如，国外许多期刊要求在提交论文时提交相关的科学数据，而我国类似科学数据开放共享的要求则较少，致使科学数据存储呈现较高的封闭性，降低了科研数据的重用性，造成了科研投入的浪费，阻滞了高水平科学数据库的建设[11]。同时，科研人员由于研究需求高度支持科研数据的开放共享，但对于数据共享存在较大顾虑[12]。例如，《中国开放数据白皮书2024》显示，中国学者对开放数据的支持度保持在高位，并呈逐年增长态势，但与此同时，学界仍普遍关注并担忧数据滥用及敏感信息泄露等潜在风险。

3.3 相关政策制度体系协同不足，政府支持力度受限

（1）相关政策制度体系协同不足。我国统一的科学数据管理相关政策颁布较晚，2018年才首次出台了国家层面专门针对科学数据管理的政策文件《科学数据管理办法》；其贯彻落实仍需长时间的积淀，且自颁布以来尚未进行过公开的更新颁布，难以适应当前人工智能等技术对于科学数据发展的要求，尚需进行不断完善。另外，由于仅存在单一的统一性指导文件，不同行业和领域中心颁布了多个数据管理标准和方法，标准之间的协同不足，导致在科学数据共享权益保护、科学数据开源开放等方面存在制度缺失，尚未建立起系统的全面的政策制度体系。

（2）相关管理协调机构和基础设施建设不足。目前，我国缺少类似欧盟EOSC的统一数据管理体系，数据归属权和管理权限存在争议，影响数据的长期保存和利用；而且与美国NSF、NIH等针对所资助项目推动的一系列数据管理计划[13]，以及针对科研智能发展所推动的批准首批35个NAIRR试点项目相比，我国尚缺少国家层面针对科研智能发展进行的科学数据平台和基础设施的建设计划。

3.4 科学数据库应用生态尚未成熟，与国外存在较大差距

近年来，我国科学数据库体系建设已取得较大进展，但科学数据库应用生态尚未成熟，跨学科、跨行业及跨国合作程度有待提升。一方面，国内科研人员

对国外的科学数据库管理系统、科学数据平台、数据处理技术及服务等存在着较高的依赖性[15]。这反映出我国科学数据库在相关数据质量、技术等方面存在较大的落后性，也体现了国内用户未养成使用自建数据库的意识和习惯，而用户缺位又影响了自建数据库的进一步完善和发展，亟待进行科学数据库生态体系的构建。另一方面，我国科研数据平台的国际影响力不足。在国际科学数据库方面，我国贡献总量有限，且存在显著的学科分布不平衡问题，致使相关领域的国际地位和国际话语权受限[14]；而且我国科学数据出境管理存在相关监管制度不健全、理念与政策法规等与国际差异较大、在权威性的国际事实标准组织中参与度和话语权不足等问题，加之缺乏相关的国际有效合作和对接机制，从而严重影响了我国科学数据国际话语权和跨国科学数据生态体系的构建[15]。以全球科学数据存储库注册库（re3data.org）收录的科学数据知识库数量为例，截至2025年2月，该平台共收录美国科学数据仓储1 191个，数量位居全球首位；相比之下，中国仅有86个被收录，其数量不足美国的1/10[16]。

4 我国加强支持科研智能数据库建设的建议

针对我国支撑科研智能发展的相关数据库体系建设面临的挑战，借鉴国际主要国家构建支撑科研智能发展的相关科学数据库建设经验，本文提出我国应建设支持科研智能的科学数据生态体系，重点从加强顶层设计和政府支持、推动开源开放和标准化管理、促进多元合作和实现与产业场景应用对接等方面提供支撑。

4.1 加强顶层设计，强化政府支持

（1）在顶层设计方面，强化科研智能数据库建设的战略性、基础性地位，完善相关机构和制度政策体系。① 设立专门的科研智能数据库，建立开放的部门机构。可由国家数据局协调国家自然科学基金委员会或科学技术部成立联合科研数据管理机构，统筹科研数据库顶层设计、标准制定、汇交共享、评估监管等工作，同时建立国家科研智能数据中心，为科研人员提供便捷的数据存取和计算支持。② 建立健全相关制度政策。将科学数据库建设纳入基础研究范畴，完善科研数据汇集、管理和开发的相关战略政策，建立健全包括科学数据共享制度、科学数据管理制度、科学数据标准及科学数据知识产权保护制度在内的科学数据库制度体系，基于保障国家科研安全的目的加强和完善战略性科研数据储备，以全面的制度和政策体系支持科学数据库生态体系的构建[17]。

（2）在资源支持方面，强化算力、资金等支持及基础设施建设。① 打造自主的高性能计算云平台。为科研数据的存储、计算和处理提供强有力的基础设施支撑，确保科研智能数据库生态体系的可持续发展。同时围绕国家数据中心构建科研、教育等多主体全链条合作共赢的生态。② 设立国家级科研智能数据库专项基金。加强资金支持，鼓励高校、科研机构和企业共同建设科研智能数据库生态体系，提高数据的全链条性、完整性和覆盖度。

4.2 建立开源开放分级机制，有针对性推动科研数据共享与合作

（1）建立健全基于科学数据特性的开源开放分级机制[18]。针对不同学科、不同研究阶段及不同性质的科研数据进行分级，注重学科约定俗成的开源规则对数据生态的促进作用，并差异性设立开源开放的水平和标准。例如，降低对过程性数据的开源开放要求，强化结果性数据的开源开放；对于基础科学领域提倡更高程度的开源开放，而对于应用科学或技术领域则需要在明确其知识产权的情况下进行开放。

（2）建立跨学科跨行业的科研数据汇交机制。建立衔接科学数据生产链条上、下游的数据生产方和数据需求方的沟通和对接机制，促进多元异构数据的可复用性和标识，提升协作效率。同时充分利用新型举

国体制优势，以政策、经费、考核等方式相融合推动同领域数据向富集区（如现有的数据中心）汇聚。

（3）建立健全开源开放共享的激励机制。① 鼓励各机构完善科学数据共享的奖励制度。针对个体上传科学数据的质量、数据和价值建立分级的奖励补贴制度，对于积极上传科学数据的科研人员给予相应的资金补贴和荣誉奖励。② 将科学数据库建设纳入科研管理体系，强调科学家在科学数据库建设中的主导性。设立专门的科研数据建设项目，将科研过程中所产生的数据质量和数量，以及相关科学数据库建设纳入评价考核标准；将科研数据上传和质量审核纳入项目考核机制，建立健全科学期刊及科研项目资助机构对所提交论文和项目所产生科研数据上传的机制安排，同时将所上传科学数据的质量作为评估论文和项目的重要标准。③ 以健全的科学数据权益保护机制和可信科学数据空间的打造，降低科研人员数据上传共享的顾虑与担忧。开发科研数据条目的全生命周期标识机制，实现单条数据的溯源；建立健全社区保护机制，强化社区数据监督。官方层面对新建立的科研智能数据库所有权和运营权进行明确，降低数据库所有者被侵权的担忧[19]。

4.3 建立标准化管理体系，确保数据库的高质量与可持续性

（1）推动科研智能数据库生态体系建设过程的标准化。① 健全科研智能数据库建设过程的标准化制度与流程。制定统一的数据标准和原则，推动与国际标准的接轨。基于国际主流标准化原则，如FAIR原则及CARE原则③等，根据学科类型、应用要求和使用阶段等进行统一的分级标准化制度设计和体系构建。同时，重视过程性数据的记录，可借鉴国防项目的数据记录机制，针对不同学科下不同实验过程类型的数据产生过程建立标准化的数据记录准则，包括记录条目、存储机制和核查机制等，完善相关数据库建设流程。② 建立动态的科研智能数据库评估机制。针对国内各机构的科学数据累积分布和管理情况进行核查，并对科研智能发展需求、受国外限制情况进行评估，确定相关科学数据库建设明确的痛点、难点，对高迫切性的数据库加大投资建设力度。

（2）推动科研智能数据库管理过程的标准化。① 构建分层次、分领域的专业科学数据管理机构体系。构建科学数据管理机构体系，由上述的中央推动性部门机构进行综合管理，同时各地方设置专业化的科学数据管理机构，实现数据库的可持续运营、质量控制和长期维护。围绕高质量数据的生产、管理、使用和激励等方面，强化专业管理机构的牵头作用，推动各学科领域数据的规范管理、标准制定等工作。② 制定科研智能数据的标准化存储格式，推动数据结构、元数据描述、数据访问接口的统一，提高数据的可操作性。完善高质量数据资源的加工机制，激励相关人员投身高质量数据生产和管理，形成开放高质量数据集，促进跨学科、跨领域的数据互操作性，加速人工智能在科学研究中的应用。

4.4 推进多元合作机制，构建全球科研智能数据库网络

（1）推动各机构内部及跨学科、跨行业合作。① 建立健全多元主体合作平台。构建整合科研机构、企业、学校与公众等多方资源、多元主体沟通协作的合作平台，畅通不同机构、不同学科和不同行业间的数据合作渠道，打破数据黑箱和壁垒限制，促进数据共享和交融。在此基础上，构建数据采集、存储、分析和应用全链条衔接的开放生态链，推动开放科学数据社区建设，实现科研智能数据库生态体系的良性构

③ CARE原则是一套以数据伦理与群体权利为核心的数据治理原则，由全球原住民数据联盟提出，强调数据活动应当以人为本、避免伤害并促进公平。其核心包括集体受益、治理与控制权、责任和伦理，关注数据“为谁所用、由谁决定、如何承担社会责任”。CARE通常与FAIR原则结合使用，用于弥补后者在社会公平和伦理层面的不足。

建。② 开展联合研究项目。通过项目合作，促进不同学科领域的研究人员合作，在共同解决跨学科问题的同时实现不同学科间的数据融合。通过合作项目和技术手段，探索数据交互的新机制。

（2）强化国际合作。① 推动全球科研数据共享标准的制定。加强与美欧国家和地区及“一带一路”共建国家合作，联合管理关键领域的科学数据库，如材料基因工程、气候科学等，增强我国在全球科技合作中的话语权和影响力。② 推动建立国际科研智能科学数据库联盟，在数据标准化、数据共享协议、伦理规范等方面达成共识，以提升我国在全球科技合作中的话语权。鼓励国内科研机构与国际领先数据库对接，使我国科研数据能够纳入全球科研生态体系，增强国际影响力。

4.5 加强科研数据建设与产业场景应用对接，提升科研智能的技术放大效应

（1）以实际应用场景为牵引，加速科研智能在重点行业落地。联合科研机构、龙头企业、行业联盟，围绕智能制造、绿色能源、生命科学、新材料等关键领域，建立产业导向的科研智能数据平台，推动数据的高效整合与行业应用，赋能新质生产力发展。例如，借鉴 The Materials Project 模式，搭建新材料领域的科研智能数据库，为新能源、半导体、高性能材料等行业提供数据支撑，推动新材料研发效率的大幅提升。

（2）推动产学研合作，共同构建科研智能数据生态体系。鼓励龙头企业和头部科研机构合作构建科研数据生态体系，激励企业与科研机构间的互联互通及工业级数据的开放与应用，以提升科研智能模型的实际适用性，如在自动驾驶、智能诊疗、气候预测等领域，以科研智能模型发展为其提供有力支持。

参考文献

1 郑令晗, 李晨珂. 面向AI4S的数据要素供给: 价值取向、路径选择与风险控制. 图书与情报, 2024, (3): 81-89.
Zheng L H, Li C K. Data element supply for AI4S: Value proposition, path choice and risk control. Library and Information, 2024, (3): 81-89. (in Chinese)

2 廖方宇, 汪洋, 曹荣强, 等. 第五科研范式下新型科研信息化基础平台架构与关键技术. 中国科学院院刊, 2024, 39 (12): 2048-2059.
Liao F Y, Wang Y, Cao R Q, et al. Architecture and key technologies of new research informatization infrastructure platform under the fifth research paradigm. Bulletin of Chinese Academy of Sciences, 2024, 39(12): 2048-2059. (in Chinese)

3 王炼. 美国联邦政府科学数据管理政策及实践. 全球科技经济瞭望, 2018, 33(7): 47-51.
Wang L. Policies and practices on scientific data management by US federal government. Global Science, Technology and Economy Outlook, 2018, 33(7): 47-51. (in Chinese)

4 刘思彤, 游玎怡, 陈光, 等. 科研机构数据管理体系构建研究——来自NIH和CNRS的经验启示. 全球科技经济瞭望, 2022, 37(6): 33-41.
Liu S T, You D Y, Chen G, et al. Research on data management system construction of research institutes: Experience from NIH and CNRS. Global Science, Technology and Economy Outlook, 2022, 37(6): 33-41. (in Chinese)

5 彭鑫, 邓仲华. 数据密集型科研环境下的科研数据管理框架研究. 数字图书馆论坛, 2017, 13(7): 61-67.
Peng X, Deng Z H. Study of research data management model under the data-intensive scientific environment. Digital Library Forum, 2017, 13(7): 61-67. (in Chinese)

6 卢祖丹. 国家科学数据共享服务平台运行绩效评价研究. 中国科技论坛, 2024, (5): 44-53.
Lu Z D. Empirical evaluation of the operational performance of national scientific data sharing service platforms. Forum on Science and Technology in China, 2024, (5): 44-53. (in Chinese)

7 李骐安, 孟宪飞, 张书华, 等. 基于FAIR原则的中国科学数据资源现状分析及启示. 数字图书馆论坛, 2023, 19(1): 50-57.
Li Q A, Meng X F, Zhang S H, et al. Status analysis and implications of scientific data resources in China based on

FAIR principles. Digital Library Forum, 2023, 19(1): 50-57. (in Chinese)

8 邢文明, 殷丹. 我国科学数据中心网络影响力评价研究. 高校图书馆工作, 2025, 45(1): 11-22.
Xing W M, Yin D. Research on the evaluation of network influence of scientific data centers in China. Library Work in Colleges and Universities, 2025, 45(1): 11-22. (in Chinese)

9 罗瑞云, 单嵩岩. 科学数据归档格式管理现状分析与启示. 北京档案, 2024, (12): 19-25.
Luo R Y, Shan S Y. Analysis of the current status of scientific data archiving format management and its implications. Beijing Archives, 2024, (12): 19-25. (in Chinese)

10 何依, 司莉, 刘莉. 我国开放科学数据平台协同治理现状调查研究. 图书馆学研究, 2024, (12): 52-62.
He Y, Si L, Liu L. Investigation and research on the current status of collaborative governance of open scientific data platforms in China. Research on Library Science, 2024, (12): 52-62. (in Chinese)

11 许丽媛, 钱力, 常志军. 科学数据汇交共享政策框架研究——以中国科学院文献情报中心为例. 图书情报工作, 2025, 69(3): 102-109.
Xu L Y, Qian L, Chang Z J. Research on the policy framework of scientific data collection and sharing: Taking the National Science Library, Chinese Academy of Sciences as an example. Library and Information Service, 2025, 69(3): 102-109. (in Chinese)

12 范昊, 郑小川, 热孜亚·艾海提. 我国科研数据开放共享政策供需匹配研究. 信息资源管理学报, 2023, 13(6): 156-165.
Fan H, Zheng X C, AihaitiReziya. Research on the match between Chinese research data open sharing policy supply and user demand. Journal of Information Resources Management, 2023, 13(6): 156-165. (in Chinese)

13 周文能, 刘云, 王刚波. 国内外科学数据管理与共享政策分析及对国家自然科学基金的启示. 中国科学基金, 2023, 37(1): 150-160.
Zhou W N, Liu Y, Wang G B. Policy analysis of domestic and foreign scientific data management and sharing and its enlightenment to National Natural Science Foundation of China. Bulletin of National Natural Science Foundation of China, 2023, 37(1): 150-160. (in Chinese)

14 张连分. 国际科学数据控制对我国数据安全的影响. 图书馆建设, 2025, (3): 80-91.
Zhang L F. Impact of international scientific data control on China's data security. Library Development, 2025, (3): 80-91. (in Chinese)

15 魏鑫, 孔丽华, 汪洋. 我国科学数据出境管理对策研究. 农业大数据学报, 2024, 6(2): 156-160.
Wei X, Kong L H, Wang Y. Research on countermeasures of scientific data exit management in China. Journal of Agricultural Big Data, 2024, 6(2): 156-160. (in Chinese)

16 陈铭. 国内外科学数据出版实践概况与发展趋势分析. 中国数字出版, 2025, 3(6): 31-39.
Chen M. Analysis of the current status and development trends in open scientific data publishing practices: Domestic and international perspectives. China Digital Publishing, 2025, 3(6): 31-39. (in Chinese)

17 顾昕. 数据知识产权在科学数据领域的适用. 科技中国, 2024, (5): 1-4.
Gu X. The applicability of data intellectual property in the field of scientific data. Scitech in China, 2024, (5): 1-4. (in Chinese)

18 孙瑜晨, 郑诗彦. 以开放共享为导向的科学数据分类分级制度建构:价值旨归与实现路径. 科技进步与对策, 2025, 42(12): 140-150.
Sun Y C, Zheng S Y. Construction of a scientific data classification and grading system oriented by Open Sharing: System value and realization strategy. Science & Technology Progress and Policy, 2025, 42(12): 140-150. (in Chinese)

19 史雅莉, 李晓璇, 严诗琴. 科学数据汇交中科研人员的行为意愿. 图书馆论坛, 2025, 45(6): 141-150.
Shi Y L, Li X X, Yan S Q. A study on the behavioral intention of researchers in scientific data submission. Library Tribune, 2025, 45(6): 141-150. (in Chinese)

Significance, challenges, and policy recommendations for strengthening database development to support AI for Science in China

CHEN Kaihua[1] LIU Hongxin[1] GUO Rui[2*]

（1 School of Public Policy and Management, University of Chinese Academy of Sciences, Beijing 100049, China;

2 School of Economics, Beijing Institute of Technology, Beijing 100081, China）

Abstract With the rapid emergence of research intelligence driven by big data and artificial intelligence (AI), high-quality, openly shared scientific databases have become a strategic focal point for scientific innovation and enhancing technological competitiveness. Major countries around the world are increasingly recognizing the foundational role of scientific databases in advancing basic research. While continuously strengthening their own scientific data infrastructure through a series of initiatives, they have simultaneously imposed restrictions and suppression on the development of AI technologies in China, including those involving research data. Against this backdrop, building an autonomous and controllable scientific data ecosystem to support research intelligence is of great significance for unlocking the potential of China's scientific data, ensuring national research security, and achieving a leap in independent innovation capacity. This study reviews international experiences in constructing databases that support research intelligence, analyzes the current status and challenges of scientific database development in China—including mismatches between the quantity and quality of data storage, insufficient openness and sharing of databases, inadequate coordination within the policy support system, and an underdeveloped database application ecosystem. Based on these findings, the study proposes strategies for strengthening China's scientific data ecosystem, including enhancing top-level design and government support, promoting open-source and standardized management, fostering diverse collaborations, and facilitating integration with industrial applications.

Keywords AI for Science (AI4S), scientific database, development autonomy, open source and open innovation, standardized management

陈凯华 中国科学院大学长聘体系特聘教授、国家前沿科技融合创新研究中心副主任。《中国科学院院刊》青年编委。主要研究领域为国家创新系统、创新管理与创新政策、创新计量学等。E-mail: chenkaihua@ucas.ac.cn

CHEN Kaihua Distinguished Professor of the long-term appointment system at University of Chinese Academy of Sciences (UCAS), and Deputy Director of Research Center for National Frontier S&T Integration and Innovation, UCAS. Youth Editorial Board Member of *Bulletin of Chinese Academy of Sciences*. His main research interests include national innovation systems, innovation management and innovation policy, innovametrics, etc. E-mail: chenkaihua@ucas.ac.cn

郭 锐 北京理工大学经济学院副教授。主要研究领域为数智化绿色化转型、复杂技术创新管理、科技追赶等。
E-mail: guorui012@bit.edu.cn

GUO Rui Associate Professor at School of Economics, Beijing Institute of Technology. His research focuses on digital- and intelligence-driven green transformation, complex technological innovation management, and technological catch-up.
E-mail: guorui012@bit.edu.cn

■责任编辑：武一男

*Corresponding author