美国国立卫生研究院AI就绪科学数据集建设实践与启示

翟军，王思慧，李昊原，李鑫
大连海事大学航运经济与管理学院，辽宁 大连 116026

摘要
为满足人工智能（artificial intelligence，AI）系统日益增长的数据需求，AI就绪（AI-Ready）数据成为当前AI发展的重点领域，得到各国政府的高度重视。其中，美国国立卫生研究院（National Institutes of Health，NIH）已走在AI-Ready科学数据集建设的前列，对其进行系统研究，可为我国高质量科学数据集建设提供一定的借鉴和参考。首先，采用网络调查、内容分析和案例研究方法，系统地介绍NIH的数据政策和资助政策、Bridge2AI旗舰AI-Ready数据集及AI就绪度框架的建设情况。进而，结合我国高质量数据集建设实际，提炼经验与启示，包括完善政策体系、构建标准体系和打造人才支撑体系。

关键词
科学数据；AI-Ready数据；高质量数据集

中图分类号：TP311.13 文献标志码：A doi：10.11959/j.issn.2096-0271.BDR26207

Practice and insight from the National Institutes of Health in building AI-Ready scientific dataset

Zhai Jun, Wang Sihui, Li Haoyuan, Li Xin
School of Maritime Economics and Management of Dalian Maritime University, Dalian 116026, China

Abstract
To meet the ever-growing data demands of artificial intelligence (AI) systems, AI-Ready data has become a key focus area in current AI development and has received high priority from governments worldwide. Among these, the National Institutes of Health (NIH) has taken the lead in building AI-Ready scientific dataset. A systematic study of its approach can provide valuable reference and guidance for the construction of high-quality scientific dataset in China. Firstly, employing web-based investigation, content analysis, and case study method, this paper systematically introduces NIH's data policy and funding policy, the flagship AI-Ready dataset of the Bridge2AI (bridge to artificial intelligence) research program, and the development of AI readiness framework. Furthermore, drawing on the practical reality of high-quality dataset construction in China, it distills beneficial experience and insights, including improving the policy system, establishing standard system, and building a talent support system.

Key words
scientific data, AI-Ready data, high-quality dataset

0 引言

随着大数据与人工智能（artificial intelligence，AI）的发展，生物医学领域已开启“数据+AI”驱动的新型科学研究范式[1]。2026年4月，斯坦福大学以人为本人工智能研究院发布《2026年人工智能指数报告》，总结全球AI的发展趋势[2]。该报告的“医学”部分指出，AI对医学研究的影响愈发突出，包括AI驱动的蛋白质研究、基于AI模型的临床推理、用于诊断的多智能体框架等。该报告同时指出，各领域AI模型的开发进程受限于数据而非架构。

为满足各类AI系统日益增长的数据需求，美国政府高度重视人工智能就绪（AI-Ready）数据的建设[3]。2025年7月出台的《赢得竞赛：美国人工智能行动计划》提出构建世界级的高质量AI-Ready科学数据集[4]；2026年3月发布的《国家人工智能立法框架》指出，应以AI-Ready格式向产业界和学术界提供联邦政府数据，用于训练AI模型和系统[5]。

我国政府同样重视高质量数据集建设。2026年3月颁布的《中华人民共和国国民经济和社会发展第十五个五年规划纲要》明确提出加快建设人工智能语料库，面向能源、交通、制造、教育、健康、金融等领域建设高质量数据集[6]。2026年6月，国家数据局印发《关于推进行业高质量数据集建设行动的实施方案》，提出实施提质增效行动，推动构建符合结构完整性、内容多样性、标注准确性、模型适配性等质量标准、满足AI-Ready的高质量数据集，降低训练推理成本，有效提升模型性能[7]。

美国国立卫生研究院（National Institutes of Health，NIH）是全球最大和最具权威的医学研究机构之一，已率先开展AI-Ready数据治理工作，并建设了一批旗舰AI-Ready数据集[8]。本文系统地分析和介绍NIH的相关政策、旗舰AI-Ready数据集和AI就绪度框架的建设情况，并提炼经验和启示，以期为我国政府部门和科研机构建设高质量科学数据集提供一定的借鉴和参考。

1 NIH支持AI-Ready数据的主要政策

1.1 数据政策

为应对大数据及数据科学对生物医学研究的新挑战，NIH于2018年发布首个《数据科学战略规划》，以确保由NIH资助的全部数据科学活动和相应产品能够符合FAIR原则，即数据可发现（findable）、可访问（accessible）、可互操作（interoperable）和可重用（reusable）[9]。该战略规划已认识到机器学习（machine learning，ML）、深度学习（deep learning，DL）等AI技术对数据创新的重要作用——可能在未来十年为生物医学研究带来革命性变化，支持相关的数据管理、分析和可视化工具的开发和推广。该战略规划的实施取得了丰硕成果，包括在云平台存储超过200 PB的生物医学数据、建设符合FAIR原则的数据仓库、资助AI与数据科学的研究、启动数据学者（data scholar）计划、完善数据管理和共享政策等[10]。

2025年12月，NIH发布新的《数据科学战略规划（2025—2030年）》，将“为软件、计算方法和AI提供新机遇”确立为五大战略目标之一，其中的首要子目标是

“通过AI促进全民健康”[10]。新版战略规划强调，AI在改善医学诊断和理解潜在的生物过程方面展现出广阔前景，但还面临着透明度、可重复性、伦理、法律和社会影响等方面的挑战，这需要为AI系统提供符合FAIR原则的AI-Ready数据。对于AI-Ready数据建设，该战略规划提出的主要措施有：①开发保障数据质量的指南和资源；②扩大数据来源，将合成临床数据应用于AI训练；③开发工具和开展培训，帮助科研人员创建AI-Ready数据集；④开发用于数据清洗、协调、集成和元数据收集的AI工具；⑤建设互联互通的数据资源生态系统，支持云平台之间数据及元数据的共享与互操作。

1.2 资助政策

1.2.1 提升数据AI/ML就绪度研究专项

生物医学研究领域虽然产生和积累了大量数据，但这些数据通常面向传统的计算和统计工具，大多数并不适用于AI/ML模型和工具。为落实数据政策，2021年4月，NIH依托数据科学战略办公室（Office of Data Science Strategy，ODSS）发布《提升NIH支持数据AI/ML就绪度专项指南》，连续3年（2021—2023年）共资助了108个科研项目（见**表1**），目标是将NIH研究生成的数据转化为AI/ML-Ready数据，并通过数据仓库、知识库等方式实现数据共享[11]。该指南对“AI/ML-Ready”的要求主要有：①符合FAIR原则对数据和元数据的要求；②具有AI/ML工具（如PyTorch和TensorFlow等）所要求的特定数据格式；③数据集规模必须足够大；④具有详细的描述文档，提供数据溯源、偏见、隐私、健康差异、伦理和应用场景等方面的信息；⑤通过与AI/ML应用的互动和反馈，提升对数据的理解、促进未来的再利用。

1.2.2 Bridge2AI研究计划

2021年5月，NIH共同基金启动“人工智能之桥”（bridge to artificial intelligence，Bridge2AI）研究计划，通过支持广泛采用AI来应对超越人类直觉的

表1 提升数据AI/ML就绪度研究专项的资助项目示例

序号	项目名称	研究机构	说明
1	语言神经影像的AI/ML-Ready	南卡罗来纳大学	将中风幸存者和健康个体的神经影像数据转化为ML-Ready的格式，用于研究大脑中的语言处理能力
2	开发AI/ML-Ready的老龄化轨迹文件	罗格斯大学	开发AI/ML-Ready数据集，预测医疗保险受益人的晚年衰老轨迹
3	提升R01下生成数据的AI/ML就绪度	塔夫茨医疗中心	生成老化研究数据的合成版本以保护数据隐私
4	改进元数据创作，提升相关数据集的AI/ML就绪度	斯坦福大学	重新设计技术，以推动AI-Ready数据的FAIR化
5	人体胰岛在早期1型糖尿病中的整合应激反应	芝加哥大学	准备复杂的组学数据集用于ML，以识别可通过微创检测手段测量的1型糖尿病早期预警信号
6	基因组医学可计算知识框架的开发与验证	国家儿童医院研究所	标准化基因组聚合数据库中的大规模群体基因组数据，以支持AI驱动的临床基因组学应用

复杂生物医学挑战，推动生物医学研究的发展[10]。第一阶段（2022—2025年）投资1.3亿美元，主要目标是：①生成全新的旗舰级生物医学与行为AI-Ready数据集；②开发数据管理的软件与标准；③创建自动化工具，加速生成符合FAIR原则且合乎伦理的数据集；④提供用于传播数据、伦理原则、工具及最佳实践的资源；⑤设立连接AI、生物医学与行为研究社区的培训项目。2026年1月，Bridge2AI研究计划进入第二阶段，在前期生成的AI-Ready数据集、工具和最佳实践的基础上，开发能够运用AI技术帮助改善健康状况的工具、设备，并提出新颖见解[12]。

不同于面向存量数据的提升NIH支持数据AI/ML就绪度研究专项，Bridge2AI研究计划专注于生成新的AI/ML-Ready数据，来应对生物医学和行为研究的重大挑战。为此，Bridge2AI出台《美国国立卫生研究院通向人工智能计划（Bridge 2AI）数据生成项目（OT2）（Data Generation Projects for the NIH Bridge to Artificial Intelligence (Bridge2AI) Program(OT2)）》作为指南，对研究工作提出具体要求，主要有：①生成新的、高价值、高质量的AI/ML-Ready数据集，以解决一个或多个生物医学和行为研究的重大问题，并推动AI/ML分析领域的发展；②鼓励项目从人类/动物受试者队列中收集多模态数据，可以被多个AI/ML模型用于理解生物和行为系统的动态本质；③数据采集过程遵守伦理框架和标准规范，尽可能采用自动化工具，确保数据和元数据的完整性，并最小化由受试人群选择偏差和批次效应引起的数据变异性；④制定详细的数据管理和共享计划，确保数据、工具和软件的大范围共享和重用[13]。此外，Bridge2AI还明确了“AI-Ready数据”的基本定义：在生物医学和临床环境中以合乎伦理的方式获取的数据，经过一定的预处理后，能够通过适当的AI/ML方法最优地用于训练、分类、预测、文本/图像生成或模拟，并且基于这些数据产生可解释的结果[14]。该定义强调AI-Ready的3个核心特征：①数据需经过处理，而非原始数据；②伦理合规贯穿整个生命周期；③结果可解释、可追溯。在此基础上，Bridge2AI标准工作组建立了“生物医学

表2 生物医学数据AI就绪度评估框架

等级	评估准则	子准则	说明
0级	FAIR化程度	0.a 可发现,0.b 可访问,0.c 可互操作,0.d可重用	对数据集的最基本要求:保存在数据仓库中,伴随基本的元数据和使用许可等
1级	溯源	1.a 透明,1.b 可追溯,1.c 可解释,1.d 关键参与者已标识	丰富的溯源信息,包括数据源、数据转换步骤及软件说明、关键参与者信息等
2级	特征描述	2.a 语义,2.b 统计信息,2.c 标准,2.d 潜在偏见来源,2.e 数据质量	全面地描述元数据,包括详细摘要、统计信息、数据字典、偏见来源、质量控制等
3级	模型前可解释性	3.a 数据描述文档,3.b 适用性,3.c可验证	以机器可读的文档(如datasheets等)提供应用案例、完整性、模型适配等信息
4级	伦理	4.a 合乎伦理获取,4.b 合乎伦理管理,4.c 合乎伦理传播,4.d 安全	符合公认的伦理原则,说明隐私保护措施、数据存储与访问的安全要求等
5级	可持续性	5.a 持久性,5.b 领域适配,5.c 良好治理,5.d 关联	确保原始未处理数据在符合隐私法规与留存规范的存档库中妥善保存,在数据采集环节之前建立数据治理模型等
6级	可计算性	6.a 标准化,6.b 计算可访问,6.c 可移植,6.d 情境化	遵循已建立的、有文档记录的标准,尽可能最大化跨计算资源的可移植性等

数据AI就绪度评估框架”（见**表2**），用于指导AI/ML-Ready数据集的实际创建工作[14]。该评估框架由7个依次递进的评估准则组成，其中FAIR化程度、溯源和特征描述聚焦数据的客观属性，模型前可解释性和伦理凸显主体责任，可持续性和可计算性强调生态适配，共同构成可操作的AI就绪度评估依据，推动生物医学数据走向AI适配的道路。

2 Bridge2AI旗舰AI-Ready数据集

2.1 旗舰数据集的基本情况

第一阶段（2022—2025年）的Bridge2AI研究计划共资助4个数据生成项目（见**表3**），总资助金额超过9 600万美元，建设4个旗舰AI-Ready数据集（见**表4**），分别对应4个重大挑战：临床护理AI/ML、功能基因组学、精准公共卫生和健康康复（糖尿病）。这些数据集在采集方法、数据模态和开放程度等方面各具特色，且都在持续迭代的过程中逐渐满足了**表2**所示的AI-Ready 7项准则[14]，代表着AI-Ready数据集建设在不同生物医学方向上的典型实践。其中，数据的获取方式从受控访问、申请访问到公开访问，体现了“尽可能开放、必要时受控”的数据共享与传播的伦理原则。

2.2 旗舰数据集的详细情况

2.2.1 CHoRUS数据集

CHoRUS项目旨在通过提供尽可能丰富的临床数据（覆盖美国监护病房床位总

表3 Bridge2AI研究计划资助的数据生成项目

序号	项目名称	重大挑战	研究机构	资助金额
1	聚焦患者的协作医院仓库统一标准(collaborative hospital repository uniting standards,CHoRUS):致力于公平的人工智能	临床护理AI/ML	麻省总医院	24 565 446美元
2	面向人工智能的细胞图谱(cell maps for artificial intelligence,CM4AI):利用细胞结构图谱构建可解释的基因组解码器	功能基因组学	加州大学圣地亚哥分校	21 026 388美元
3	作为健康生物标志物的声音(Bridge2AI-Voice):构建一个来源合乎伦理的生物声学数据库,以前所未有的方式理解疾病	精准公共卫生	南佛罗里达大学	17 995 278美元
4	糖尿病人工智能就绪与探索性图谱项目(artificial intelligence ready and exploratory atlas for diabetes insights,AI-READI):揭示人类健康如何在疾病后恢复的细节,以2型糖尿病(type 2 diabetes mellitus,T2DM)为模型	健康康复(糖尿病)	华盛顿大学	32 682 088美元

表4 Bridge2AI旗舰AI-Ready数据集的基本情况

序号	数据集名称	发布平台	获取方式	数据规模
1	CHoRUS	MGB Azure	受控访问	超过23 TB
2	CM4AI	LibraData	公开访问	21.4 TB
3	Bridge2AI-Voice	PhysioNet	申请访问	不详
4	AI-READI	FAIRhub	申请访问	3.82 TB

数的2%以上），推动重症监护领域的AI/ML研究。CHoRUS数据集汇集了美国14家医院的真实临床数据，涵盖加强监护病房（intensive care unit，ICU）、儿科重症监护治疗病房（pediatric intensive care unit，PICU）和新生儿重症监护治疗病房（neonatal intensive care unit，NICU）的入院患者数据。截至2026年4月，数据集已含有2.6万名入院患者的数据（最终将超过10万名）、16亿行标准的电子健康记录（electronic health records，EHR）、7 642例放射学影像数据和23 TB的波形遥测数据[15]。数据集中的数据模态达到9种，如结构化数据（如人口统计、诊断、用药等）、半结构化数据（如临床笔记）、非结构化数据（如影像数据）和连续型生理信号数据（如波形数据）等。

项目通过多模态数据标准化处理框架（见**图1**），对不同类型的数据采用差异化的标准化技术，如处理结构化数据的观察性医学结果伙伴关系（observational medical outcomes partnership，OMOP）通用数据模型、处理半结构化和非结构化文本文档的开放健康自然语言处理（open health natural language processing，OHNLP）工具包、处理影像数据的医学数字成像和通信（digital imaging and communications in medicine，DICOM）标准和存储波形遥测数据的波形数据库（waveform database，WFDB）格式，进而在统一的云平台上实现数据关联。目前，CHoRUS完整数据集处于受控访问状态，用户必须接入CHoRUS云安全区、签署许可协议后方可使用。

由于涉及大量的个人信息，该项目成立了专门的伦理小组，识别、分析与应对CHoRUS数据集在AI/ML应用中的伦理、法律与社会问题。项目将伦理、法律审查安排在各个阶段的转换时间点上，逐步开放数据访问权限：第一阶段只支持本地访问，医院之间数据不互通；第二阶段仅限CHoRUS项目内部访问，数据被划分为训练数据集和验证数据集；第三阶段允许公众授权访问训练数据集，前提是通过NIH的研究人员认证；第四阶段公开发布训练数据集。

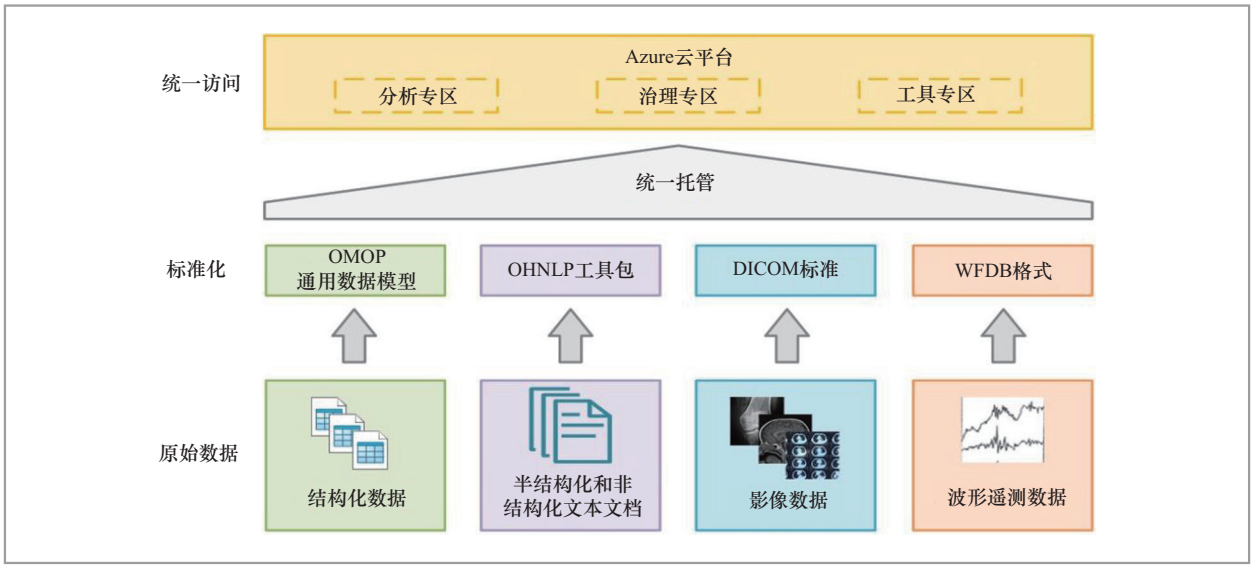

图1 CHoRUS多模态数据标准化处理框架

2.2.2 CM4AI数据集

CM4AI数据生成项目绘制人类细胞的时空结构图谱，整合细胞成像、蛋白质组质谱及成簇规律间隔短回文重复序列（clustered regularly interspaced short palindromic repeats，CRISPR）基因扰动3种多模态数据，构建可解释的基因型-表型学习模型[16]。该项目直面当前基因组学AI的“黑箱”困境，通过将AI模型的输入锚定在明确的生物学机制上，实现模型前（pre-model）的可解释性。

细胞图谱（cell maps）是层次化的有向无环图，其中每个结点代表在特定尺度上相互靠近的蛋白质组装体[17]。构建细胞图谱的数据流来源于一系列受到扰动和未受扰动的细胞系，包括已处理和未处理的癌细胞系，以及分化的和初始的诱导多能干细胞。CM4AI采用自动化的生产流程生成AI-Ready数据（见图2），包括数据获取、数据集成、AI-Ready封装和数据出版4个主要阶段。其中，前两个阶段将数据及其元数据打包成研究对象封装包（research object crate，RO-Crate）格式，再由项目开发的框架对数据、软件、元数据和溯源图进行AI-Ready封装，最后将RO-Crate数据包（RO-Crate package）发布到NIH批准的综合仓库平台上。

目前，CM4AI数据集的规模已达21.4 TB，涉及的蛋白质超过7 000种、靶向基因超过11 000个，采集的免疫荧光图像超过53 000份。CM4AI数据集的开放版本发布在弗吉尼亚大学LibraData平

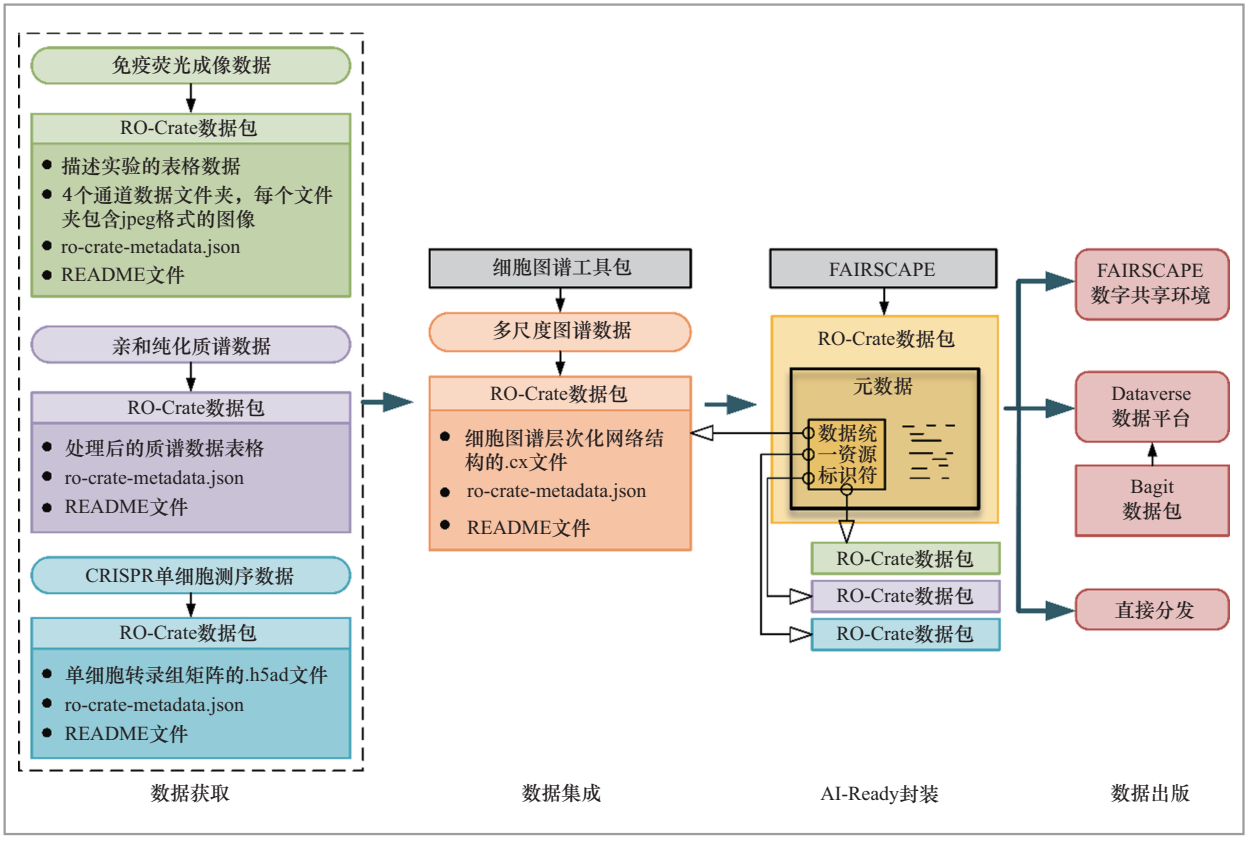

图2 CM4AI AI-Ready数据的生产流程

台上，包含7个数据子集和1个元数据文件包，在知识共享-署名-非商业性使用-相同方式共享4.0（CC BY-NC-SA 4.0）协议许可下供研究人员非商业性免费使用，商业用途需要与版权持有者（加州大学圣地亚哥分校、斯坦福大学、加州大学旧金山分校等）进行单独的许可协商[17]。

2.2.3 Bridge2AI-Voice数据集

人类语音构成了与疾病状态相关的丰富信息来源，凭借其多样的声学特性，加上良好的成本效益和可及性，语音作为疾病筛查、诊断和监测潜在生物标志物的效用已获得认可[18]。将AI/ML与语音分析相结合，有望在健康诊断、筛查和监测中发现可扩展的声学标志物，从而改善患者预后。在此背景下，Bridge2AI-Voice项目的目标是建立一个大规模、符合伦理来源的、与健康元数据相关联的语音、言语和咳嗽数据库，以支持AI驱动的研究。该项目研发了一款新型智能手机应用程序——Bridge2AI-Voice App，用于收集声学任务的标准化录音、经过验证的患者问卷和患者报告结果。在数据采集环节，项目组制定了声学标准，严格规范麦克风类型、录制环境、采样率及校准流程，确保各采集站点之间语音数据的一致性、可比性与科学有效性[19]。在数据预处理环节，项目组开发了工具包Prepkit，将语音与问卷数据转换为脑成像数据结构（brain imaging data structure，BIDS）兼容格式，同时完成元数据清洗、重组与特征转换等工作。

Bridge2AI-Voice数据集发布在生物医学研究平台PhysioNet上，该研究平台处于受限访问状态，用户需取得资质并签署数据使用协议后才能下载数据。其中，成年人数据集（3.0版）涵盖北美5个站点833名成年人的33 041次录音（157.5 h）数据，儿童数据集（1.0版）涵盖300名儿童（2~18岁）的22 620次录音（24.45 h）数据。值得注意的是，平台发布的并非原始音频，而是衍生的语音特征数据，这一设计平衡了参与者的隐私保护和数据分析价值。

为吸引更多的研究人员关注和使用Bridge2AI-Voice数据集，Bridge2AI-Voice联盟已连续3年（2023—2025年）主办“Voice AI研讨会”。2025年的研讨会标志着语音生物标志物科学从概念研究向临床实施的转变，2026年研讨会的主题是“跨越桥梁：从回声到洞见，再到可持续的影响力——探索语音AI在医疗保健领域的过去、现在与未来”。

2.2.4 AI-READI数据集

AI-READI是一项针对T2DM的数据采集项目，旨在促进AI/ML在健康生成（从T2DM向健康复原力的转变）研究上的应用[20]。AI-READI项目分为6个模块，分别是数据采集、伦理保障、标准建设、团队协作、工具开发以及技能与劳动力发展[21]。数据采集方面，该项目将从3个地点招募4 000名40岁及以上人群，按糖尿病严重程度（无糖尿病、糖尿病前期、口服药物控制、胰岛素治疗）分成4组（每组各1 000人）。各个地点遵循相同的研究方案，采集每位参与者的多种类型数据，包括调查反馈、视力评估结果、视网膜成像、心电图、持续血糖监测、居家环境测量和活动监测数据，还计划提取EHR。

各个地点的数据均映射到适用的标准格式，并通过严格的质量控制，定期上传到AI-READI专用数据管理平台FAIRhub，实现数据的共享和AI-Ready。该平台支持即时的AI/ML研究，不需要进

行数据的重新格式化和预处理。AI-READI数据集当前版本为v3.0.0，包含2023年7月19日至2025年5月1日期间采集的2 280名参与者的多模态数据，不含受保护的健康信息（protected health information，PHI），参与者的性别、种族/族裔以及所用药物信息也被删除。AI-READI数据集的数据文件总数为356 343个，数据规模达到3.82 TB。下载数据需先验证身份，且仅将数据用于T2DM的相关研究，并接受“AI-READI数据使用许可”。

为提升AI就绪度水平，AI-READI项目采用自动化工具评估FAIRhub平台上数据集的“FAIR化程度”。同时，为优化数据集的AI就绪度，项目特别研制了“临床数据集结构（clinical dataset structure，CDS）”标准规范，用于组织和描述临床研究数据集，使其易于互操作及人类和机器的重用。

3 FAIRSCAPE——Bridge2AI的AI就绪度框架

3.1 FAIRSCAPE框架的功能

FAIRSCAPE是由弗吉尼亚大学医学院用Python语言开发的计算框架，用于保障科学数据遵循FAIR原则和AI-Ready准则。FAIRSCAPE框架最初应用于CM4AI项目，随后被推广至其他Bridge2AI项目，成为落实AI就绪度准则的有效支撑工具。

FAIRSCAPE框架的主要功能是数据集生成AI-Ready所需要的各种元数据，为数据和软件提供符合FAIR原则的持久标识符，并与Harvard Dataverse软件（NIH批准的仓库平台）无缝集成，支持AI-Ready数据集的高效出版和在线发布[22]。

3.2 FAIRSCAPE框架的核心技术

（1）RO-Crate元数据封装。RO-Crate是由欧盟Horizon 2020、德国研究联合体和哈佛大学共同资助的社区项目，提供FAIR研究对象（FAIR research object，FRO）及其元数据封装的技术方案，已被广泛应用于生物信息学和数字人文等领域[23]。FAIRSCAPE框架采用最新的RO-Crate元数据规范1.2版，为数据集提供名为ro-crate-metadata.json的元数据文件，主要描述两类实体：①数据实体（data entity）代表实际的文件或目录，类型为file或dataset，通过hasPart/isPartOf属性建立数据集之间的关联，元数据项有名称（name）、编码格式（encodingFormat）、内容大小（contentSize）、许可（license）等；②上下文实体（contextual entity）代表描述数据所需的外部事物，如人员（person）、机构（organization）、地点（place）等，通过@id属性相互引用。

CM4AI数据集的RO-Crate元数据文件可从LibraData平台下载，其中的部分元数据记录如**表5**所示。

（2）溯源描述与可视化。“深度溯源”（deep provenance）是AI-Ready数据区别于普通FAIR数据的本质特征之一，这要求数据能够通过所有中间步骤（包括软件、模型和计算）回溯到其未修改的原始来源[14]。FAIRSCAPE框架采用“证据图本体（evidence graph ontology，EVI）”对数据集的溯源信息进行语义建模。EVI本体模型由弗吉尼亚大学开发，包含37个类77个属性。其中，所有实体类（entity）子类均可解析到归档的数字组件，并可使用领域本体术语进行标注。这一细粒度的子类化使生物医学数据溯源能够精确区分不同类型的实体与活动，满足

表5 CM4AI数据集的部分RO-Crate元数据记录

序号	元数据项	元数据值	说明
1	conformsTo	https://w3id.org/ro/crate/[redacted]	遵循RO-Crate规范1.2版
2	name	Cell Maps for Artificial Intelligence	数据集的名称
3	@type	Dataset	类型为数据集对象
4	keywords	Cell maps, AI, CM4AI, Machine learning, et al	关键词集合
5	license	https://creativecommons.org/licenses/[redacted]	数据集的使用许可
6	contentSize	19.1 TB	数据规模
7	ethicalReview	Vardit Ravistky ravitskyv@thehastingscenter.org	伦理方面的联系人
8	identifier	https://doi.org/10.18130/V3/[redacted]	唯一标识符
9	hasPart	@id: https://fairscape.net/view/[redacted] rocrate-untreated-if-data-release et al	数据集包含的数据子集
10	publisher	@type: sc:Organization name: https://dataverse.lib.[redacted]	数据集的发布者,类型为机构
11	ro-crate-metadata	Perturb-Seq/cell-atlas/ro-crate-metadata.json	RO-Crate元数据文件的URL

了AI-Ready准则1.b（可追溯）关于“确定关键的数据转换步骤”的要求。

同时，EVI引入EvidenceGraph（证据图）类，表达由数字对象及其支持/质疑关系构成的有向无环图。FAIRSCAPE框架支持AI-Ready数据集EVI元数据记录和可视化EvidenceGraph的生成，CM4AI数据集一个子集的EvidenceGraph实例如**图3**所示，其中展示了数据处理所需要的数据集（dataset）、计算（computation）和软件（software）等数字对象。

（3）Croissant元数据集成。Croissant是由MLCommons社区主导开发的专门描述ML-Ready数据集的元数据规范，已应用到OpenML、Hugging Face和Kaggle等主流的ML平台上[24]。为弥补RO-Crate元数据对AI/ML特定需求覆盖不足的局限，FAIRSCAPE选择性地集成了Croissant RAI（responsible AI）扩展版中的部分元数据元素，包括数据偏见和

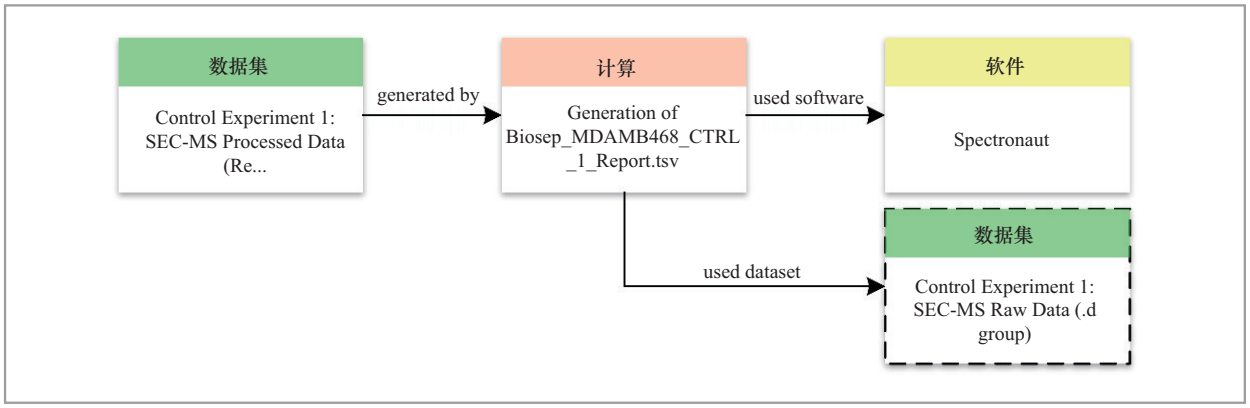

图3 EvidenceGraph实例

局限说明、数据使用案例等。这些元数据直接映射到AI-Ready准则2.d（潜在偏见来源）“说明数据已知的偏见来源”的要求，及准则3.b（适用性）“明确数据集适用与不适用场景”的要求。

FAIRSCAPE框架为数据集提供名为ro-crate-croissant.json的元数据文件，**表6**给出了CM4AI数据集的部分Croissant元数据记录。

（4）数据模式的自动验证。为保障数据集的结构一致性和质量（准则2.c和2.e），FAIRSCAPE集成了Frictionless Data技术框架[25]。该技术框架包含一套开源工具和规范，提供数据模式推断、定义和验证等功能[26]。FAIRSCAPE客户端将Frictionless Data框架与RO-Crate技术紧密集成，当用户上传数据时，客户端会预先执行模式验证功能。这一机制将数据质量检查前移至客户端，避免将无效数据上传到服务器，大幅度提高了数据处理的效率与可靠性。

通过封装和实现RO-Crate、EVI、Croissant和Frictionless Data等标准规范，FAIRSCAPE框架将抽象的AI就绪度准则（见**表2**）映射到具体的技术实现（见**表7**），将数据集的AI/ML-Ready保障工作转化为易于操作的自动化工作流[14]。

3.3 FAIRSCAPE框架的结构

FAIRSCAPE框架的结构如**图4**所示，包括客户端和服务器端两部分。客户端通过命令行接口（command line interface，CLI）和图形用户界面（graphical user interface，GUI）等方式，帮助用户打包和验证元数据组件——RO-Crate数据包；服务器端的应用程序编程接口（application programming interface，API）接收传入的RO-Crate数据包，提取、处理和验证元数据，为每个RO-Crate数据包、软件、数据集和计算对象分配持久标识符，随后将数据和元数据存

表6 CM4AI数据集的部分Croissant元数据记录

序号	元数据项	元数据值	说明
1	conformsTo	http://mlcommons.org/croissant/[redacted]	遵循Croissant RAI规范
2	dataReleaseMaintenancePlan	Dataset will be regularly updated and augmented on a quarterly basis …	数据维护计划：数据将定期更新和扩充
3	personalSensitiveInformation	Involves human subjects: None – data collected from commercially available cell lines …	个人敏感信息：不涉及人类受试者，数据来源于商业可获得的细胞系
4	dataBiases	Data does not represent all biological variants which may be seen in the population at large …	数据偏见说明：不代表在普通人群中可能观察到的所有生物学变异
5	dataLimitations	This is an interim release. It does not contain predicted cell maps, which will be added in future releases …	数据局限说明：不包含预测的细胞图谱，将在未来版本中添加
6	dataUseCases	A major goal is to enable biologically-driven, interpretable ML applications, for example as proposed in Ma et al. 2018 (PMID: 29505029) …	数据使用案例
7	dataCollectionTimeframe	9/1/2022, 10/13/25	数据采集的时间范围

表7 FAIRSCAPE框架封装的技术标准

序号	技术标准	作用	支持的AI就绪度准则
1	RO-Crate	将数据与元数据打包为可独立分发的单元，支持附加型、分离型两种模式	准则0(FAIR化程度) 准则6(可计算性)
2	EVI	构建深度溯源元数据，记录数据、软件、计算活动与代理之间的派生关系	准则1(溯源) 准则3(模型前可解释性)
3	Croissant	提供AI/ML任务专用元数据(数据分割、特征定义、偏见标注等)，兼容ML生态	准则2(特征描述) 准则3.b(适用性)
4	Frictionless Data	提供数据模式的推断、定义与验证，确保数据结构一致性	准则2.c(标准) 准则2.e(数据质量)

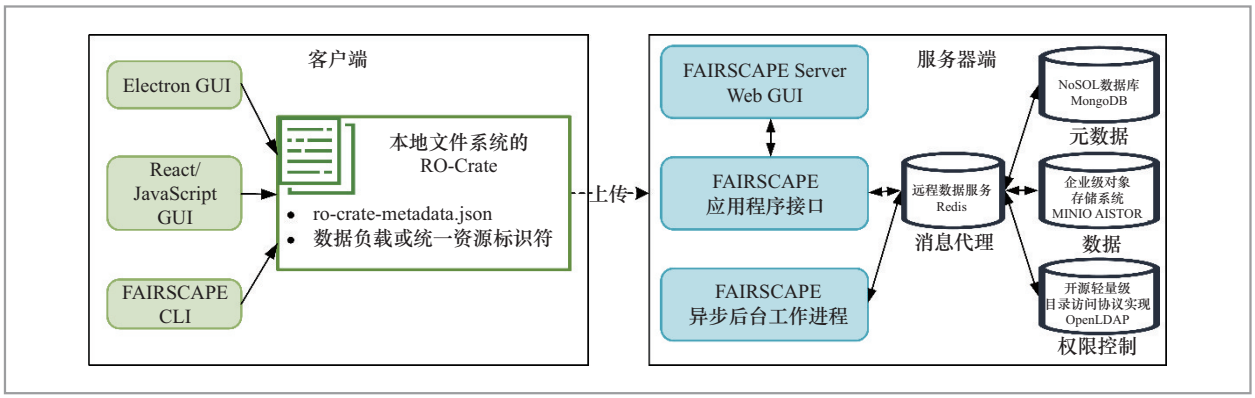

图4 FAIRSCAPE框架的结构

储到数据库中。

FAIRSCAPE框架是开源的，其源代码发布在GitHub，采用麻省理工学院(Massachusetts Institute of Technology，MIT)许可证供用户下载和使用，体现了Bridge2AI项目对软件FAIR原则的支持。

4 经验与启示

2025年8月，国家数据局征集遴选出104个高质量数据集典型案例，涵盖科学研究、工业制造、智慧能源、交通运输、医疗卫生、教育教学等12个重点领域，以及低空经济、具身智能、智能驾驶、智慧海洋、生物制造5个创新领域[27]。截至2026年3月，我国已建成高质量数据集11.6万多个，数据总规模超过960 PB。2026年4月，国家数据集管理服务系统上线运行，发布数据集1 200余个，标志着我国高质量数据集建设工作迈入集约化管理新阶段。同时，AI正从大语言模型向多模态模型、从基础模型向行业模型、从内容生成向智能体自主决策、从数字智能向具身智能演进，对数据集的供给规模、质量和AI就绪等提出了更高的要求。

总体而言，我国的高质量数据集建设仍处于起步和探索阶段，存在着优质资源

供给缺口较大、开放共享程度偏低、AI-Ready不足、标准制定和能力建设滞后等问题[28-29]，需要借鉴和吸收美国等国家的先进做法和成熟经验。

4.1 完善政策体系

在政策层面，NIH的数据政策和AI政策协同一致，合力推动AI-Ready数据集的规范管理和开放共享，这一点也体现在NIH的上级管理机构——美国卫生与公众服务部（Department Of Health & Human Services，HHS）的相关政策上。《HHS数据战略（2023—2028年）》将培养数据人才、促进数据共享与负责任地使用AI列为优先领域，并提供资金资助提升AI/ML就绪度的研究[30]；《HHS AI战略计划》明确支持FAIR原则、开放数据和开源软件原则，以加速推进AI-Ready数据及标准、Bridge2AI旗舰数据集的建设[31]。

建设高质量数据集同样是我国数据战略和AI战略的共同关注点之一，但在打破数据孤岛、实现跨部门跨地区数据共享的政策支持上，我国与欧美等地区和国家之间还存在明显差距[27]。2018年3月，国务院办公厅印发《科学数据管理办法》，提出要规范科学数据管理，保障数据安全、提高开放共享水平[32]。作为国家层面的科学数据政策，《科学数据管理办法》已支持国际上通用的部分FAIR数据原则，但在建设高质量数据集的背景下，仍有较大的优化空间，主要有：①不仅支持FAIR数据原则，还支持FAIR软件原则，推动数据、软件与工具的协同开放共享；②重视AI数据供给，引入AI-Ready理念和准则；③融入AI数据治理的最新成果，包括质量管理、安全管理、元数据管理和伦理管理等，提升AI数据的规范化管理水平。

4.2 构建标准系统

建设AI-Ready数据集，Bridge2AI重视发挥标准规范体系的引领和保障作用[25]。Bridge2AI联盟成立专门的跨越项目边界的标准工作组，捕捉社区需求、汇集标准要求与规范，协调标准的实施、共享与复用。为此，Bridge2AI联盟建设了标准注册和查询系统，收集标准782项，其中生物医学标准（320项）、数据和元数据标准（97项）、本体/词汇表（179项）和软件/工具（144项）等。

Bridge2AI标准工作组秉承“开放共享”原则，与AI-Ready数据生成同步，搭建起一套开放的可推广、可复用的标准规范体系，主要特点是：①在FAIR原则和AI就绪度准则的指导下，选择和研制相关标准，如可复现神经影像计算中心（acenter for reproducible neuroimaging computation，ReproNim）主导开发的开源标准ReproSchema、自研的临床数据结构（clinical dataset structure，CDS）和声学标准（Acoustic Standards）等；②在研制新标准的同时，大量复用研究社区共同遵循的开放标准（见**表8**）；③在各个旗舰数据集之间共享标准，如DICOM、OMOP CDM等（见**图5**）。

2026年4月1日，全国数据标准化技术委员会发布《高质量数据集 格式要求》《高质量数据集 分类指南》和《高质量数据集 质量评测规范》等国家标准的征求意见稿，标志着我国高质量数据集的标准建设已步入快车道。借鉴NIH经验，除了研制新标准，我国还应积极引进和吸收国际上普遍采用的开放标准规范，如已被多个AI社区使用的Croissant元数据规范等，以提升数据资源的国际化程度和国际影响力。重要的是，还应支持和鼓励科研团队在建设高质量数据集的同时，搭建起相应

表8 Bridge2AI项目复用的开放标准举例

序号	标准名称	研究社区
1	通用数据模型(CDM)	观察性医疗结果合作组织(OMOP)
2	Open mHealth	开放移动健康研究社区
3	RO-Crate技术规范	RO-Crate 社区
4	Croissant元数据规范	MLCommons社区
5	脑成像数据结构(BIDS)	BIDS开放全球社区

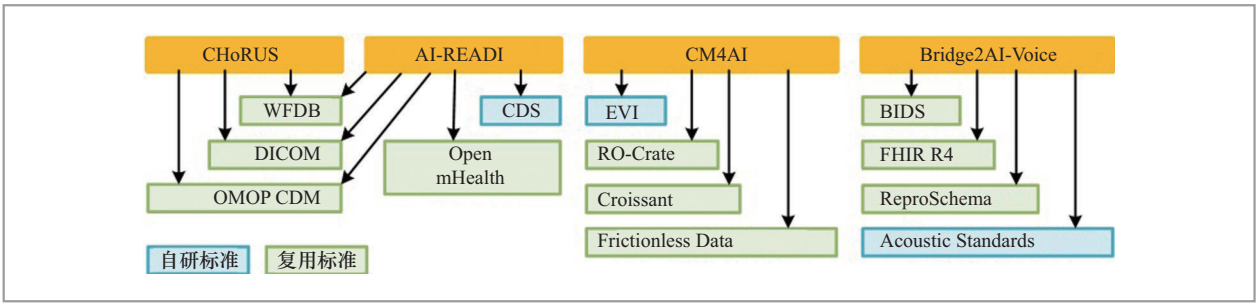

图5 Bridge2AI各个数据集遵循的数据和元数据标准

的标准规范体系，并将该体系推广到研究社区或行业，通过标准化促进数据集在各个AI领域的整合、集成与互操作。

4.3 打造人才支撑体系

由国家数据局指导编制的《高质量数据集建设指引》指出，“加强人才培养”是形成高质量数据集可持续发展产业生态的重要一环，包括引导高等院校、职业院校加强学科建设，开设数据治理、数据分析等专业课程，为高质量数据集建设提供充足的人才储备。深化产教融合，鼓励校企合作建设实训基地，提高在校学生的实践技能，建设高素质的人才队伍。吸引数据科学家、工程师等行业专家参与重点领域高质量数据集建设，培养具有专业背景的高端人才。完善国家数据标注等职业认证体系，开展专项培训，规范数据标注人才发展，提高从业者的数据治理与跨领域协调能力[27]。

具体到科学数据集建设领域，NIH通过各个层面的多项举措致力于人才培养和人才梯队建设，有效提升了科研人员的数据素养和AI能力。

（1）NIH层面的举措。为应对具有重大公共健康影响的生物医学数据的挑战，NIH于2020年启动数据学者计划，招募AI、云计算、数据科学、数据库管理、软件设计、超级计算和生物信息学等领域的科学家及工程师，参与或主持NIH的数据项目和AI项目。2020—2022年，已吸引20多名学者加入NIH的不同研究所，包括国家眼科研究所、国家癌症研究所、国家生物医学影像与生物工程研究所等。NIH还于2021年启动了“信息科学、AI/ML与生物医学交叉领域的人才队伍建设”项

目，培养精通FAIR数据和AI-Ready数据相关技能的新型研究人员。

(2) Bridge2AI层面的举措。Bridge2AI联盟成立“培训、招聘与指导”(training, recruitment and mentoring, TRM) 工作组，开展生物医学/行为学研究与AI研究的跨学科协作式培训工作。在Bridge2AI计划的研究基础上，TRM工作组开发教育模块及实践活动，并探讨与之交织的社会与伦理议题；建设教育资源平台，覆盖全链条AI/ML人才培养体系的教学资源，包括伦理、法律与社会影响、AI基础与生成式AI、大语言模型及前沿技术等。由NIH全额资助的在线课程融合生物医学真实场景，由顶级专家加盟授课，不仅面向研究人员和医生，还面向在校本科生和研究生。

(3) 项目层面的举措。CHoRUS项目依托丰富的数据资源，面向临床医生、数据科学家和数据科学学生，开设培训课程并同步举办“数据马拉松”竞赛；CM4AI团队与Kaggle合作，举办CM4AI生物网络图社区检测竞赛，吸引参赛者拓展AI/ML在生命科学领域的应用边界；Bridge2AI-Voice项目面向全球各地的学生，举办“Voice AI暑期学校”，培养使用Bridge2AI-Voice数据集的“Bridge2AI-Voice学者”，为未来的研究人员、开发者和科学家储备人才；AI-READI团队则为研究生、博士后和实习医生等提供为期一年的实习项目，旨在扩充数据科学/AI与生物医学及临床交叉研究领域的人才储备。

5 结束语

Gartner发布的2025年人工智能技术成熟度曲线表明，AI-Ready数据是当前AI发展的热点领域之一。中美两国无疑是这一领域的领先者，中国重视场景应用和行业数据，而美国更重视“AI驱动的科学研究”(AI for Science, AI4S) 和科学数据。

生物医学是AI4S的重点领域，该领域的AI-Ready数据建设具有标杆示范意义。本文聚焦美国NIH的AI-Ready科学数据集建设，采用内容分析、案例研究等方法，分析和介绍其政策、旗舰AI-Ready数据集和AI就绪度框架，总结政策、标准和人才方面的实践经验及启示。显然，我国也应高度重视生物医学研究领域的AI-Ready高质量数据集建设，加大资金和人才投入力度，并开发该领域的AI-Ready准则、评估工具、标准规范和保障系统，加快推动AI在辅助诊疗、健康管理和医保服务等场景的应用，以大幅提高医疗健康服务能力和效率。

后续研究将持续关注Bridge2AI旗舰数据集的开发应用情况，进一步探索AI-Ready数据在应用中所面临的一系列问题，如可解释性、溯源、伦理和安全等。

参考文献：

[1] 邹丽雪. 美国推动人工智能在生物医学领域应用的政策分析及其对中国的启示[J]. 科技管理研究, 2023, 43(16): 31-38.
Zou L X. Policy analysis of promoting the application of artificial intelligence in the biomedical field in the United States and its implications for China[J]. Science and Technology Management Research, 2023, 43(16): 31-38.

[2] HAI. Artificial intelligence index report 2026 [R]. 2026.

[3] 钱力, 刘志博, 胡懋地, 等. AI就绪的科技情报数据资源建设模式研究[J]. 农业图书情报学报, 2024, 36(3): 32-45.
Qian L, Liu Z B, Hu M D, et al. Construction model of AI-ready for scientific and technological intelligence data resources[J]. Journal of Library and Information Science in Agriculture, 2024, 36(3): 32-45.

[4] The White House. Winning the race: AMERICA'S AI Action Plan[EB]. 2025.

[5] The White House. A national policy framework for artificial intelligence[EB]. 2026.

[6] 中华人民共和国国民经济和社会发展第十五个五年规划纲要[EB]. 2026.
Outline of the 15th five-year plan for national economic and social development of the People's Republic of China [EB]. 2026.

[7] 国家数据局. 国家数据局关于印发《关于推进行业高质量数据集建设行动的实施方案》[EB]. 2026.
National Data Administration. Circular of the National Data Administration on issuing the implementation plan for promoting the development of high-quality industry datasets[EB]. 2026.

[8] 王茜, 刘湘闽, 陈凌云, 等. 美国国立卫生研究院AI就绪数据治理实践与启示[J]. 医学信息学杂志, 2025, 46(7): 1-6, 14.
Wang Q, Liu X M, Chen L Y, et al. Practice and implication of AI-ready data governance of the U. S. national institutes of health[J]. Journal of Medical Informatics, 2025, 46(7): 1-6, 14.

[9] NIH. NIH strategic plan for data science[EB]. 2018.

[10] NIH. NIH strategic plan for data science, 2025-2030[EB]. 2025.

[11] NIH. About the administrative supplements to support collaborations to improve the AI/ML-readiness of NIH-supported data[EB]. 2025.

[12] NIH. Bridge to artificial intelligence (Bridge2AI)[EB]. 2026.

[13] NIH. Data generation projects for the NIH bridge to artificial intelligence (Bridge2AI) program (OT2) [EB]. 2021.

[14] Clark T, Caufield H, Parker J A, et al. AI-readiness for biomedical data: Bridge2AI recommendations[J]. bioRxiv, 2024: 2024.10.23.619844.

[15] Davidson A E, Jose A, Shickel B, et al. Building AI competence in the healthcare workforce with the AI for clinical care workshop: a Bridge2AI for clinical CHoRUS project[J]. Journal of Clinical and Translational Science, 2025, 9(1): e243.

[16] CM4AI Project. Cell maps for artificial intelligence[EB].2026.

[17] Clark T, Mohan J, Schaffer L, et al. Cell maps for artificial intelligence: AI-ready maps of human cell architecture from disease-relevant cell lines[J]. bioRxiv, 2024: 2024.05.21.589311.

[18] Moothedan E, Boyer M, Watts S, et al. The Bridge2AI-voice application: initial feasibility study of voice data acquisition through mobile health[J]. Frontiers in Digital Health, 2025, 7: 1514971.

[19] Bridge2AI-Voice Project. Bridge2AI-voice tools & resources [EB]. 2026.

[20] Owsley C, Matthies D S, McGwin G, et al. Cross-sectional design and protocol for artificial intelligence ready and equitable atlas for diabetes insights (AI-

READI) [J]. BMJ Open, 2025, 15(2): e097449.

[21] AI-READI Project. Artificial intelligence ready and exploratory atlas for diabetes insights [EB]. 2026.

[22] Al Manir S, Levinson M A, Niestroy J, et al. The FAIRSCAPE AI-readiness framework for biomedical research [PP]. bioRxiv [preprint]. 2026: 2024.12.23.629818.

[23] Soiland-Reyes S, Sefton P, Crosas M, et al. Packaging research artefacts with RO-crate[J]. Data Science, 2022, 5(2): 97-138.

[24] Akhtar M, Benjelloun O, Conforti C, et al. Croissant: a metadata format for ML-ready datasets[PP]. V3. arXiv (2024-12-09) [2026-05-26]. arXiv: 2403.19546.

[25] Caufield H, Ghosh S, Kong S W, et al. Standards in the preparation of biomedical research metadata: a Bridge2AI perspective[PP]. V2. arXiv (2025-09-16) [2026-05-26]. arXiv: 2509.10432.

[26] Fowler D, Barratt J, Walsh P. Frictionless data: making research data quality visible[J]. International Journal of Digital Curation, 2017, 12(2): 274-285.

[27] 国家数据局. 高质量数据集建设指引 [EB]. 2025.
National Data Administration. Release of guidelines for high-quality dataset construction[EB]. 2025.

[28] 姜春宇, 白玉真, 刘渊, 等. 构建企业级人工智能高质量数据集: 方法与路径[J]. 大数据, 2025, 11(6): 47-56.
Jiang C Y, Bai Y Z, Liu Y, et al. Building high-quality datasets for enterprise-level artificial intelligence: methods and pathways[J]. Big Data Research, 2025, 11(6): 47-56.

[29] 杨琳, 朱扬勇. 高质量数据集产品的形态和生产流程研究[J]. 大数据, 2026, 12(2): 64-74.
Yang L, Zhu Y Y. Research on the form and production process of high-quality dataset products[J]. Big Data Research, 2026, 12(2): 64-74.

[30] HHS Office of the Chief Data Officer. 2023-2028 HHS data strategy [EB]. 2023.

[31] HHS artificial intelligence strategy & implementation [EB]. 2025.

[32] 邢文明, 肖嘉丽, 陈继丽. 面向FAIR原则的《科学数据管理办法》优化研究[J]. 图书馆论坛, 2022, 42(4): 86-95.
Xing W M, Xiao J L, Chen J L. Research on the optimization of measures for the management of scientific data under the FAIR principles[J]. Library Tribune, 2022, 42(4): 86-95.

作者简介



翟军（1969- ），男，博士，大连海事大学航运经济与管理学院教授，主要研究方向为信息管理、开放数据、数据治理。



王思慧（2001– ），女，大连海事大学航运经济与管理学院硕士生，主要研究方向为数据质量管理。



李昊原（1999– ），男，大连海事大学航运经济与管理学院硕士生，主要研究方向为元数据管理。



李鑫（2002– ），女，大连海事大学航运经济与管理学院硕士生，主要研究方向为数据溯源管理。

录用日期: 2026-06-29

通信作者: 翟军，zhaijun@dlmu.edu.cn

基金项目: 教育部人文社会科学研究规划基金项目（No.25YJAZH029）

Foundation Item: The Humanities and Social Sciences Planning Fund of Ministry of Education (No.25YJAZH029)