

基础学科高质量科学数据集建设与应用实践研究

高瑜蔚^{1,2}, 朱艳华^{1,2}, 胡良霖^{1,2}, 赵欢^{1,2}, 郑培森^{1,2}, 李亚杰^{1,2}

1. 中国科学院计算机网络信息中心, 北京 100083;

2. 国家基础学科公共科学数据中心, 北京 100083

摘要

在AI for Science驱动全球科研范式变革的背景下, 高质量科学数据集已成为支撑基础学科创新的核心生产要素, 其建设质量直接关系到我国基础研究核心竞争力。针对当前科学数据资源分散、标准不统一、质量参差不齐、成果转化不畅等行业痛点, 国家基础学科公共科学数据中心立足多学科发展需求, 探索形成多维度协同的高质量数据集建设体系。本文结合中心实践成果, 系统阐释基础科学高质量数据集的核心内涵, 提出培育权威数据库、遴选AI4S数据集、形成用户高评价数据集、建设数据论文预印本平台、升级智能体服务体系等五大实践路径, 剖析数据应用价值与转化机制, 可为我国基础科学数据集规范化建设、开放共享及AI-ready成果培育提供理论参考与实践借鉴。

关键词

基础科学; 高质量数据集; AI for Science; 数据治理; 开放共享

中图分类号: G203

文献标志码: A

doi:10.11959/j.

issn.2096-0271.2026xxx

Exploration of High-Quality Dataset Development and Application Practices in Basic Disciplines

Gao Yuwei^{1,2}, Zhu Yanhua^{1,2}, Hu Lianglin^{1,2}, Zhao Huan^{1,2}, Zheng Peisen^{1,2}, Li Yajie^{1,2}

1. Computer Network Information Center, Chinese Academy of Sciences Beijing 100083, China;

2. National Basic Science Data Center Beijing 100083, China

Abstract

Driven by the global research paradigm transformation of AI for Science, high-quality scientific datasets have become core productive resources for the innovative development of basic disciplines, fundamentally determining the competitiveness of China's basic research. To address the prevalent issues of fragmented data resources, inconsistent standards, uneven data quality, and inefficient transformation of research outcomes, the National Basic Science Data Center has constructed a systematic, multidimensional framework for high-quality dataset development. Based on the center's

practical experience, this paper expounds the core connotation and construction criteria of basic science datasets and summarizes five key practices: the development of authoritative databases, the construction of AI for Science-oriented dedicated datasets, the integration of user-feedback-optimized high-quality datasets, the establishment of a scientific data preprint platform, and the deployment of an AI-based intelligent service system. This study further analyzes the application value and transformation mechanism of scientific data. The findings provide theoretical support and practical references for the standardized development, open sharing, and AI-ready output of basic science datasets in China.

Key words

basic science, high-quality datasets, AI for Science, data governance, open sharing

1 引言

当前,人工智能与基础科学的深度融合催生了人工智能驱动科学发现(AI for Science)这一新型科研范式,推动科研模式从传统“假设驱动”向现代“数据驱动”加速转型。在日趋激烈的国际科技竞争背景下,科学数据作为科研创新的核心要素,是支撑基础学科实现跨越式发展的关键载体;其质量优劣、可用程度与共享水平,将直接影响AI模型训练效果、科学发现效率及高水平科研成果产出质量。因此,研究基础学科高质量科学数据集建设路径具有紧迫性和必要性。

2 科学数据集建设现状和特点

国际上,各国在政策层面日益重视科学数据集工作,特别是美国2025–2026年密集出台顶层政策,先后发布《赢得竞赛:美国人工智能行动计划》《国家人工智能政策框架:立法建议》《科学:一个新的黄金时代》报告等,将高质量科学数据集的建设与开放利用提升至国家核心竞争力的高度,形成了覆盖战略规划、立法保障、落地执行的完整政策链条。实践层面,欧美

等主要国家与地区已形成较为系统的高质量科学数据集建设与应用体系,欧洲开放科学云(EOSC)以FAIR原则为治理框架推动科研数据共享复用,美国国家科学基金会(NSF)通过国家级数据基础设施建设支持高质量数据集研发,美国国立卫生研究院(NIH)以FAIR原则为核心推进AI-ready数据建设。国际实践正朝着以高质量数据集建设为核心、标准化治理为基础、跨域应用为导向的一体化方向发展。

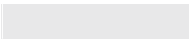
我国高度重视科学数据的管理与应用工作,各相关部门协同发力,先后出台一系列政策文件,明确数据要素与高质量科学数据集的建设要求、发展方向与标准规范,形成上下联动、协同推进的工作格局。2022年12月,国务院印发《中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见》,为高质量科学数据集建设提供顶层制度指引;2024年12月,国家发展和改革委员会联合相关部门发布《关于促进数据产业高质量发展的指导意见》与《国家数据基础设施建设指引》,明确支持基础科学等领域打造高质量数据集;国家数据局持续推进高质量数据集标准体系建设^[9],明确数据质量核心要求,推动数据“供得出、流得动、用得好、保安全”,为基础科学高质量数据集建设提供全方位政策与标准支撑;科技部出台《科学数据管理办法》与《“十四五”国家基础

研究规划》，推动科学数据全流程规范化管理与开放共享；国内学者围绕高质量数据集建设开展多维度研究，在质量评估方法^[1-2]、评价体系^[3-7]、行业应用^[8-11]、AI驱动治理^[12-14]、科学数据领域实践^[15-19]等方向形成系统成果，但在基础学科场景下的体系化构建、跨学科融合及AI-ready持续优化路径方面的研究仍需深化；数据孤岛、标准不一、质量不均、数据安全、数据伦理等问题仍制约科研应用效果^[20-25]。

依托多部门政策统筹推进，我国基础科学数据体系建设已取得阶段性成果，但受科研创新活动特殊性等因素制约，仍存在不少瓶颈问题，影响了数据要素赋能基础科研能力。其一，数据资源布局碎片化，跨机构、跨学科数据壁垒难以打通，数据孤岛普遍存在。其二，细分领域数据标准不完善，现有标准执行力不足，数据全流程管理缺乏统一规范。其三，全域数据质量管控体系缺失，噪声干扰等问题频发，无法有效支撑AI模型训练、前沿基础研究等科研需求。其四，数据治理与人工智能技术融合不足，数据要素价值转化链路存在堵点，数据赋能科研的价值未能充分释放。结合真实科研场景看，上述瓶颈具有鲜明的具象化特征。比如，国家重点研发计划项目所产生的科学数据面向特定科技创新需求，定制化特征突出，研究场景单一、样本规模有限，跨项目与跨领域复用难度大；野生动植物野外观测数据受自然环境与物种属性限制，存在小众物种样本断层、野外数据噪声大、专业标注成本高昂等问题；活体人脑成像数据存在类似短板，数据获取成本高昂，受设备运维、伦理审查、受试者招募等多重约束，有效样本体量不足；同时扫描伪影与人脑个体先天差异会产生固有噪声，常规自动化标注

工具无法适配，必须依托专业神经科学研究人员开展人工标注。综上，基础学科数据集具有样本稀缺、应用门槛高、噪声来源复杂等特征，通用数据治理方案无法匹配其专业化的数据整合、质控与智能标注需求。

国家基础学科公共科学数据中心（以下简称“国家基础数据中心”）作为我国基础学科领域科学数据汇聚、治理、共享与服务的国家级核心基础设施，长期致力于构建具有国际影响力的权威科学数据库，探索高质量数据集标准化建设范式，以完整性、权威性、战略性、稀缺性为建设准则，推动数据资源高效转化为AI-ready研究支撑能力^[26]。相较于同领域其他数据库与数据服务平台，国家基础数据中心深耕物理、化学、植物、动物等基础学科及前沿交叉技术领域，数据资源主要来源于科研人员产生的原创性科学数据以及国家重点研发计划等国家级重大科研项目汇交的数据，此类体系完整、标准规范、权威性强的基础科学数据资源，在国内同类平台中具有突出的独特性与稀缺性。与此同时，国家基础数据中心突破了多数同类平台学科门类单一、数据体系碎片化的局限，现已覆盖60类一级学科、357类二级学科，实现了基础学科全域立体化覆盖，可有效支撑基础学科前沿探索与AI for Science领域的科研创新工作。国家基础数据中心通过“面向科研需求汇聚优质数据、聚焦AI for Science建设专用数据库、依托用户反馈优化数据质量、创新数据出版模式、升级门户智能服务”五大核心路径，构建多学科高质量数据资源体系，实现部分数据集的商业价值转化。本文立足我国基础科学数据发展现状，结合国家基础数据中心的实践探索，系统阐述基础学科领域高质量数据集建设标准、路径与应



用实践，为基础学科数据驱动型的科研创新提供实践指引。

3 高质量科学数据集内涵与路径

3.1 核心内涵

本文认为，基础学科高质量数据集是指面向数学、物理、化学、生物、天文、地球科学等基础学科领域，经过标准化采集、清洗、结构化、标注、质检与合规化处理，能够支撑 AI 模型训练、算法验证、科学发现及跨学科研究需求，具备高可用性、高可靠性、高安全性与高适配性的数据集合。其核心价值在于为基础科学研究提供“精准数据燃料”，有效解决科研数据质量不佳、可用性不足的痛点，同时支撑 AI for Science 新型范式落地应用，推动基础科学领域的原创性发现与技术突破。

与普通数据集相比，基础学科高质量数据集具备三大典型特征：一是权威性，数据权威来源规范，经领域专家严格审核与多轮校验，能够客观反映领域研究的前沿水平与核心成果；二是规范性，数据全生命周期严格遵循统一的数据采集、存储、标注与元数据标准，确保数据的一致性与互操作性；三是适配性，紧密结合基础学科研究特点与 AI 技术应用需求，完成数据结构化、特征提取等预处理，可直接用于 AI 模型训练与科研实践。

3.2 建设路径

立足基础科学研究实际需求并对标 FAIR 原则（可发现、可访问、可互操作、可重用），

国家基础数据中心结合长期工作实践，形成了“源头管控—过程治理—质量校验

—开放共享”的数据全生命周期建设路径，实现了数据集从采集、治理到应用、转化的全过程高质量保障。这一路径以全流程可控、全要素可溯、全场景可用为目标，将质量要求嵌入数据生命周期的每一环节。具体包括：

（1）源头规范采集是保障数据质量的首要前提。明确各学科数据采集范围、技术方法与实施流程，优先遴选权威科研机构、重大科技项目产出的高质量数据，并根据学科特性制定差异化采集方案，同步采集完整元数据以记录来源、采集时间、处理方法与环境参数，从源头减少数据缺失、噪声与偏差。（2）全流程精细治理是提升数据规范性与 AI 适配性的关键。建立覆盖清洗、标注、结构化转换、合规化处理的一体化治理体系，融合领域知识与智能算法进行处理，通过去重、补全、精准标注、非结构化数据转换及脱敏溯源，使数据集更适配科研分析与 AI 模型训练。（3）质量校验为数据集的科学性与可用性提供可靠保障。采用 AI 自动校验与领域专家审核相结合的双重核验机制，对关键指标进行批量检测与问题整改，并由专家审定科学价值与权威性，同时建立动态监测与定期复核机制，实现数据质量持续优化。（4）开放共享与价值转化以“开放优先、安全可控”为原则。通过标准化平台为数据集赋予科技资源标识符（CSTR），面向科研与产业主体提供分级服务，并建立用户反馈闭环。

在此基础上，国家基础数据中心进一步构建“数据—科研—成果”一体化价值转化体系，推动高水平论文发表、专利技术转化与产业应用落地，打破数据孤岛，支撑跨学科研究与原创科学发现，最终实现高质量数据集在基础科研、AI 研发与产业创新中的多元价值释放。

4 具体实践

立足基础学科创新发展需求，对标国际一流数据基础设施建设标准，国家基础数据中心以“权威数据库引领、重点课题驱动、用户反馈优化、平台创新赋能、智能服务升级”为多元驱动路径，系统构建覆盖多领域、多尺度、多模态的高质量科学数据集体系，已形成一套可复制、可推广的科学数据治理与服务模式。

4.1 培育国际权威科学数据库

国家基础数据中心依托国内顶尖科研机构的力量，建设并长期运行维护一批具有国际影响力的基础科学权威数据库，严格按照国际通用标准进行数据汇交、质量管控、更新维护与开放共享服务，成为我国基础科学领域数据走向世界的重要窗口，同时也为生物、植物、核物理、脑科学等领域 AI-ready 研究与科学发现提供了关键数据支撑。

（1）《中国生物物种名录》由中国科学院动物研究所主导，联合 300 余位国内外领域专家共同编制，严格遵循国际物种 2000 名录规范，是我国生物多样性领域最具权威性的核心数据集。该名录自 2008 年起持续进行年度更新，系统整合中国动物、植物、菌物等全类群基础数据，2025 年依据最新文献新增动物物种 3000 种，并同步优化在线查询功能、数据导出及接口服务，有效支撑了重点科研任务实施，为物种自动识别、空间分布预测、生物多样性评估等 AI 模型训练提供了高质量、高可信度的标注数据。

（2）世界山茶属植物品种数据库是目

前全球数据规模最大、标准体系最严格的山茶属专题数据库，由国家基础数据中心联合科研机构共建，全面收录全球各国茶花、茶叶、茶油品种信息，包含已灭绝品种，并已整合品种数据 5.7 万余个，其中有效接受名 2.7 万个，配套图片资源 6.9 万余张。数据库秉承国际开放共享原则，95% 的图片资源对全球用户开放，获得国际山茶协会认证并被国际园艺学会收录为权威数据来源，同时为上海植物园创建国际杰出茶花园提供了鉴定与普查支撑，直接服务于园艺学、植物学领域品种分类、生长预测等研究。

（3）核物理主题数据库由中国原子能科学研究院承建，系统汇聚国内外权威核科学数据资源，集成中国评价核数据库（CENDL）、国际核反应实验数据库（EXFOR）、衰变数据库（CENDL-DDL-1.0）等核心库，内容覆盖中子诱发核反应、核素衰变方式、半衰期、射线能量与强度等关键指标。数据库建立了严格的质量控制流程，公开版本已全面开放共享，广泛应用于核能工程、核安全、资源勘探、生物医学等应用，为核反应过程模拟、辐射剂量预测、核设施安全评估等 AI 模型研发提供了稳定可靠的数据基础。

（4）中国活体人脑成像共享数据库由中国科学院心理研究所承建，是国际上具有重要影响力的脑影像公共数据平台，专注于活体脑成像数据与认知行为数据的标准化整合与共享。目前已收录 6~18 岁健康青少年多模态脑影像及配套行为数据 102 例，联合科学数据银行共建“彩巢计划-中国人脑智毕生发展数据社区”，推动跨学科数据融合与复用。数据累计下载量超过 44 万次，支撑了多项高水平研究在 Nature 子刊 Scientific Data 等期刊发表，已成为神

经科学、认知心理学及类脑 AI 研究不可或缺的高质量数据资源。

4.2 建设 AI for Science 高质量数据库

国家基础数据中心面向 AI for Science 前沿发展需求，以重点课题遴选、年度专项建设任务落地为实施路径，锚定 AI 模型训练与前沿科学问题挖掘核心诉求，构建专用高质量科学数据库。传统科学数据库的设计逻辑以文献归档、单实体静态存储、人工检索服务为核心，而 AI 专用高质量数据库则采用多模态结构化组织方式、批量流式存储架构及机器可读标准化交互接口，二者在数据处理范式上存在本质差异；同时，适配 AI 计算场景的就绪数据集，还需要满足多模态统一对齐、细粒度结构化标注、批量高效读取及溯源元数据描述等系列技术规范。

在长期实践过程中，国家基础数据中心探索形成“质量指标引领、领域知识赋能、技术工具支撑”的专用数据库建设模式：该模式深度融合基础学科领域知识与前沿信息技术，将专业知识与规范化流程、AI 技术工具有机结合，系统性提升数据库建设效率与质量，推动高质量数据资源的协同发展。

2024 年，国家基础数据中心在高质量数据库建设中，重点遴选并推进 5 个 AI for Science 专用高质量数据库的建设，分别为中国动物声音标准数据库、空间生物学行为分析数据库、化学物质环境风险数据库、高质量天然产物科学数据库和多模态机器学习数据库。这些数据库紧密围绕 AI 模型训练需求，在建设过程中严格遵循统一的数据标准与规范，重点提升数据的标注精度、特征适配性与多样性，针对性地解决了相关领域 AI 研究面临的数据短

缺、质量不佳等痛点问题。

中国动物声音标准数据库聚焦动物声学研究 with AI 识别的核心需求，规范动物声音数据的采集、标注与存储流程，系统收集不同物种、不同场景下的动物声音样本，构建标准化的动物声音特征库，为动物物种识别、行为分析等 AI 模型训练提供数据支撑；空间生物学行为分析数据库围绕空间生命科学研究需求，收集空间环境下模式生物（线虫、果蝇、斑马鱼等）的行为数据，进行标准化标注与结构化处理，支撑空间生物学领域 AI 驱动的行为模式分析与机制研究；化学物质环境风险数据库整合各类化学物质的环境暴露数据、毒性数据，构建风险评估数据集，为 AI 驱动的环境风险预测、污染物溯源等研究提供数据保障；高质量天然产物科学数据库汇聚天然产物的结构、活性、提取工艺等相关数据，规范数据标注与分类体系，支撑 AI 驱动的天然产物筛选与药物研发研究；多模态机器学习数据库聚焦多学科多模态数据整合需求，汇集文本、图像、数值等多类型数据，完成模态对齐与特征提取，为跨模态 AI 模型训练提供基础数据支撑。

4.3 基于用户反馈形成高评价数据集

国家基础数据中心坚持“需求导向、用户至上”的建设理念，以国家科技项目数据汇交与治理服务为重要抓手，深化科学数据集的服务与应用，通过建立完善的用户评价体系，系统收集科研团队、企业及行业机构的反馈意见，实现数据集质量的持续优化，形成一批用户认可度高、应用价值突出的高评价数据集。

在国家科技项目数据汇交服务中，国家基础数据中心严格落实科技项目数据汇交的相关政策，对各类国家重点研发计划、

科技重大专项等项目汇交的数据进行规范化治理与专业化服务，明确数据汇交标准、流程与质量要求，确保汇交数据的合规性与可用性。同时，搭建用户反馈平台，建立“数据服务—反馈收集—问题整改—质量提升”的闭环机制，通过线上问卷、线下座谈、一对一咨询等方式，广泛收集用户在数据集使用过程中提出的问题与建议，重点关注数据准确性、完整性、适用性及服务便捷性等核心需求。

针对收集到的用户反馈，国家基础数据中心首先对反馈内容进行结构化整理，并按数据内容、元数据描述、数据服务和AI应用适配性等类型归类，形成数据质量优化清单。对于数据值错误、标注不准确等问题，重点进行数据核验、标注复核和重新标注；对于字段缺失、说明文件不完整等问题，重点开展字段补全、说明文件完善和补充采集；对于格式不统一、单位不一致、命名不规范等问题，重点开展格式规范化、单位统一和元数据修正。通过上述处理，用户反馈从分散意见转化为明确的数据治理需求和服务优化方向，为后续数据集质量提升提供依据。在此基础上，国家基础数据中心组建专业治理团队，对用户反馈集中、应用需求明确的数据集开展针对性优化，持续完善数据内容、元数据描述、说明文件和服务功能，提升数据集的可理解性、可获取性与AI应用适配性。同时，国家基础数据中心持续提升数据资源的服务效能与知识价值，通过在线调查、邮件反馈等多种渠道，广泛收集用户评价，评选出广受用户认可的高价值数据集，并对这些数据集的内在逻辑关联进行深入分析，形成一系列主题明确的高评价数据集，进一步强化数据资源的应用价值与共享效能。

以“中速磁浮列车性能与运行测试专

题数据集”为例，该数据集是国家基础数据中心于2024年基于用户反馈评价较高整合构建的高质量科学数据资源，内容涵盖中速磁浮列车集成试验最高速度（200公里/小时）、运控系统控车运行状态、牵引效率、运控牵引半实物集成仿真测试、悬浮能耗及最大爬坡能力仿真等关键维度；数据全部来源于国家重点研发计划项目，具有高度的自主性与权威性。华东交通大学江西省先进控制与优化重点实验室在使用后对该数据集给予高度评价，认为其在“有用性”方面深化了对中速磁浮测试与半实物仿真框架的理解，对数据驱动建模具有参考价值与共享再利用潜力；在“准确性”方面数据无误、表述清晰；在“完整性”方面内容与说明文件一致，信息丰富全面。

通过持续的用户反馈与优化迭代，国家基础数据中心已形成一批高评价、高质量科学数据集。这些数据集在科研应用中表现突出，用户满意度达95%以上，进一步提升了数据资源的应用效能，为AI-ready研究提供了更贴合实际需求的数据支撑。

4.4 科学数据预印本提升数据集成果属性

国家基础数据中心紧扣科学数据开放共享与数据要素价值释放的重大需求，针对传统学术出版周期冗长、数据资产难以即时确权、数据汇交质量不高等现实瓶颈，提出科学数据出版与项目数据汇交深度融合的新型服务模式，自主研发并上线我国首个科学数据预印本发布平台——DarXiv，并于2024年10月正式投入运行，推动科技计划项目汇交数据的质量提升与开放共享应用。该平台支持数据论文及配套科学数据集的快速发布与共享，推

动使科学数据集形成的数据论文能够像学术论文一样发布并获得认可，激励数据质量的持续提升。平台通过规范化元数据描述体系与数据学术社区建设，提升了科学数据的可发现性、可理解性与可复用性，成为科学数据出版流通机制与科技项目数据汇交治理模式的重要创新实践。

在科技项目数据汇交管理中，国家基础数据中心严格遵循国家科学数据管理相关政策，将科学数据预印本发布机制深度嵌入科研项目全生命周期数据管理体系。传统汇交模式通常以项目结题验收为节点，数据汇交多表现为合规导向的被动式归档，难以匹配科研节奏，造成高价值数据资源发布滞后、复用率偏低等问题。DarXiv 平台通过引入科学数据预印本制度，使科研团队在数据产生阶段即具备数据首发权，支持科研人员在项目实施过程中，将阶段性形成的可共享数据集及配套数据论文及时提交并发布。平台对上线数据论文统一分配科技资源标识符（CSTR），并结合版本信息、变更记录和审核日志，对数据论文及配套数据集的发布、更新和引用过程进行管理，实现数据资产的唯一标识、动态溯源与有效确权。当数据内容、元数据描述或说明文档发生更新时，平台保留相应版本记录，支持科研人员在引用和复用数据时定位到具体版本，从而提升科学数据出版、引用和再利用过程中的可追溯性与可复现性。该模式提升了科研人员的数据共享意愿，推动数据治理节点从事后被动汇交向事前规划、事中主动出版转变，形成“数据出版促进项目汇交、项目汇交丰富出版资源”的良性生态，为科学数据高效汇聚、规范治理与开放共享提供了制度创新支撑。

截至2026年5月，DarXiv科学数据预印本平台已累计发布数据论文3500余篇，

服务国家重点研发计划项目400余项，联合国内外科研机构320余家，数据资源覆盖电子与通信技术、材料科学、人工智能、光学工程、公共管理等多个前沿交叉学科领域。为保障数据资产的科学性、规范性与合规性，平台建立了严格的内部审核机制，所有数据论文及配套数据集均需经项目负责人或课题负责人审核通过后方可发布，确保数据集逻辑严谨、元数据规范完备、样本说明清晰准确，符合所属领域的科学标准与质控要求。对后续更新的数据资源，平台同步记录更新原因、更新内容、审核结果及发布时间，确保数据版本变化过程可追踪、可核验。

依托严格质控与开放共享机制，DarXiv平台为数据密集型科学研究与AI for Science领域提供经过科研共同体初步筛选的高质量、可直接复用的数据资源。应用实践表明，数据集配套的数据论文对采集环境、预处理流程、算法设计及参数体系进行了系统阐述，从而降低了后续研究人员的数据理解成本与复用门槛，在数据驱动建模、算法验证与开放共享再利用方面具有重要的参考与应用价值。

4.5 AI智能服务强化数据集服务效能

国家基础数据中心坚持以用户需求为核心导向，围绕高质量数据集“找得到、找得准、用得好”的建设目标，积极推进科学数据服务模式向智能化、语义化、场景化转型，有效解决传统关键词检索与目录浏览难以适配科研实际需求的痛点，破解用户“有数据找不到、找到数据不会用、不确定是否适用”的现实难题。为进一步强化高质量数据集对AI for Science研究的支撑能力，推动数据资源向AI-ready成果高效转化，2025年，以门户服务智能

化升级为重点，完成门户服务智能体的整体设计与原型研发，构建了数据集智能推荐助手与智能问答助手两大 AI 服务模块，全面提升数据服务的便捷性、精准度与智能化水平。

在技术实现上，现阶段智能服务体系采用通用大模型 API、RAG 检索增强和 MCP 高级检索接口相结合的技术路线。系统围绕数据集元数据、数据论文全文和业务问答文档构建知识库。用户输入自然语言需求后，首先进行意图识别与关键词扩展，随后分别通过知识库检索和 MCP 高级检索接口召回候选数据集。其中，知识库检索主要用于获取数据集的语义描述、数据论文上下文信息及推荐依据，MCP 高级检索接口则主要用于调用门户的检索能力，获取与用户需求匹配的数据集列表及相关结构化信息。系统对两路召回结果进行字段标准化、合并去重和相关性排序，最终生成数据集推荐表单、推荐理由和可追溯依据。

(1) 数据集智能推荐助手以自然语言交互为入口，面向科研人员提出的研究主题、数据对象、应用场景等需求开展资源发现。对于模糊化、口语化的查询需求，系统通过关键词扩展和语义匹配提升候选数据集召回能力；对于多来源的召回结果，系统结合数据集元数据、数据论文内容和门户检索结果进行综合排序，优先推送与用户需求匹配度高、数据说明充分、应用价值明确的数据集，并以表单形式展示推荐结果。该方式能够在一定程度上弥补传统关键词检索依赖精确词匹配、难以理解科研语义需求的不足，降低用户发现和筛选数据资源的成本。

(2) 智能问答 AI 助手面向数据服务全流程，聚焦数据汇交、在线申请、使用规范、权限管理、AI 适配方法等高频场景，

构建标准化问答知识库，通过自然语言处理技术实现自动应答、流程指引与业务咨询，有效降低人工服务压力，提升用户自助服务能力与整体服务效率。以 2026 年 1 月 1 日至 2026 年 5 月 31 日基于智能问答对话记录进行初步统计分析，结果显示智能问答助手累计形成对话 644 个，服务用户 594 人，用户提问 1117 次，AI 回答 1117 次，其中有效回答 1028 次，占回答总数的 92.03%。总体上，智能问答助手能够覆盖多数常见数据服务咨询需求，在数据申请、汇交流程、使用规范、权限管理等高频问题中形成较稳定的自动响应能力，有效支撑了人工咨询压力的降低与用户自助服务效率的提升。

通过上述智能服务体系建设，中心形成了“用户需求—智能发现—推荐反馈—质量提升”的闭环运行机制，持续提升高质量数据集的可发现性、可理解性与可利用性，不断优化数据服务能力，为 AI-ready 数据集建设与基础学科科研创新提供全流程、智能化、高效率的数据支撑保障。

5 挑战与展望

5.1 面临挑战

尽管国家基础数据中心在高质量数据集建设、开放共享与场景化应用方面取得了一定的成效，但面向 AI for Science 深度融合、跨学科创新与数据要素市场化配置的更高要求，基础科学高质量数据集建设仍面临一系列现实挑战。

第一，跨学科数据整合难度较大，不同基础学科在数据标准、结构格式及质量规范上存在明显差异，多源异构数据难以实现高效互通、关联融合与统一治理，制

约了跨学科研究与大模型训练的数据支撑能力。第二，数据治理智能化水平有待提高，数据标注、质量校验、异常识别、动态更新等关键环节仍以人工干预为主，自动化、智能化治理工具覆盖有限，导致大规模数据集建设效率偏低、一致性保障难度较高。第三，数据安全、伦理合规与开放共享之间的平衡问题突出，生物、医学、核科学等领域敏感数据在开放共享中面临隐私泄露、知识产权及安全管控等风险，安全可控前提下的高效共享机制仍需完善。第四，数据成果转化机制尚不健全，科学数据向高水平科研成果、技术专利及产业应用的转化路径受阻，产学研协同衔接不足，数据要素价值难以充分发挥；同时，我国数据出版规范与相关标准体系的国际影响力有限，全球话语权仍有待提升。

5.2 未来展望

未来，面向基础科学原创突破与数据强国建设战略目标，国家基础数据中心将坚持问题导向与需求牵引，持续完善高质量数据集建设体系，推动数据治理能力向智能化、融合化、可信化、国际化、产业化升级，为 AI for Science 科研范式提供更强支撑。

一是全面推进数据治理智能化。深化人工智能技术在数据采集、标注、质检、更新、溯源中的全流程应用，构建智能审核、自动标引、知识图谱融合、质量动态监控等智能化治理体系，提升大规模数据集建设效率与标准化水平，降低人工依赖与治理成本。二是强化跨学科数据融合与标准统一。建立跨领域元数据互通机制与数据融合框架，推动学科领域数据标准协同与接口兼容，打破机构与学科壁垒，构建多模态、跨尺度、高质量的基础科学融

合数据集，支撑大模型训练与交叉科学研究。三是构建安全可控的开放共享与可信数据空间。完善数据分级分类、脱敏审查与权限管理体系，运用隐私计算、联邦学习、区块链等技术实现敏感数据“可用不可见”；依托已获批的“行业可信数据空间创新发展试点（基础科学）”，构建覆盖实名认证、质量评级、价值评估、授权管理、争议解决的全链条运营机制，形成数据供给、使用、服务、运营、监管多方协同的可信流通生态。四是完善成果转化与国际标准协同发展体系。深化产学研多方合作，搭建成果转化对接平台，打通数据向科研论文、技术专利、算法模型及产业应用的转化链路，构建数据与科研创新互促迭代的良性循环，充分释放数据要素价值。同时，持续完善数据出版、质量评价、元数据等标准体系，积极参与全球科学数据治理，逐步提升我国在相关规则与标准制定中的国际影响力与话语权。

6 结束语

基础学科高质量数据集是支撑 AI for Science 发展、推动科研范式转型的重要基础，其建设水平直接关系到我国基础学科的创新能力和国际竞争力。国家基础数据中心通过“国际权威数据库引领、重点课题驱动、用户反馈优化、平台创新赋能、智能服务升级”的多元驱动模式，构建了覆盖多个学科的高质量数据集体系，形成了标准化、规范化的建设路径，实现了数据资源的开放共享与多元价值释放。同时，其在预印本平台协同发布、智能服务门户体系构建等方面的创新实践，进一步拓展了高质量数据集的建设范式，完善了数据服务模式。

本文以国家基础数据中心的建设实践为依托，系统阐释了基础学科高质量数据集的核心内涵与应用路径，深入剖析了当前我国在数据集建设、治理与共享过程中面临的现实挑战，并展望了未来发展方向。研究表明，高质量数据集的构建必须遵循标准引领、质量优先、需求导向、开放共享的核心理念，唯有通过技术创新赋能、制度体系完善和跨领域协同推进，才能充分释放科学数据作为核心生产要素的关键价值。展望未来，随着数据治理智能化水平的持续提升、成果转化机制不断健全以及国际合作深度拓展，基础学科高质量数据集将为我国基础科学研究与人工智能技术深度融合提供坚实支撑，进而推动我国在 AI for Science 全球竞争中构筑先发优势，助力基础学科实现高水平创新与跨越式发展。

参考文献：

[1] 李继峰,张成龙,刘鑫,等.面向人工智能的数据治理框架[J].大数据,2025,11(1):3-20.

[2] 罗春旭,熊海旭,叶雅珍,等.TDQE:一种面向深度学习的文本数据质量评估方法[J].大数据,2025,11(06):95-107.

[3] 王锋,沙若男,张智超,等.高质量数据集建设运营能力成熟度评估体系研析[J].质量与认证,2025, No.230(12): 41-43.

[4] 杜方伟,宋吉述.出版业高质量数据集建设评价指标与运营策略[J].中国出版,2025, No.609 (16): 3-10.

[5] 林镇阳,胡鑫,郭明军,等.可信数据空间协同治理下的高质量数据集建设与长效运营路径[J].图书情报知识,2025, 42(05): 19-30.

[6] 高瑜蔚,方缙,胡良霖,等.面向科学数据集社会评论的细粒度情感分析方法[J].重庆邮电大学学报(自然科学版),2025,37(4):607-616.

[7] 高瑜蔚,王兆钧,胡良霖,等.面向科学数据集决策影响力评价的长文本分类方法[J].北京邮电

大学学报,2025,48(2):8-17.

[8] 王玉珏,许瑞婷,焦俊杰,等.档案文献遗产高质量数据集建设路径研究[J].北京档案,2026, 421(01): 7-16.

[9] 徐拥军,张群群,傅予,等.哲学社会科学高质量数据集的核心特征、应用需求与建设进路[J].图书情报知识,2025,42(06): 6-15+27.

[10] 罗梦杰,赵云鹏,张梦轩,等.大模型时代的石油化工行业高质量数据集:挑战与机遇[J].化工学报,2026, 77(01): 435-447+538.

[11] 张博睿,陈桃,胡婕,等.企业高质量数据集建构:内涵特征、逻辑框架与未来展望[J].信息资源管理学报,2026.

[12] 张成岗,夏子叶.大模型迭代下高质量数据集建设:技术演进、现实挑战与策略趋向[J].江苏行政学院学报,2026, No.146(02): 129-139.

[13] 张群洪,李林,杨堃,等.基于可信数据空间架构的高质量数据集建设[J].大数据,2026.

[14] 梁正.数据要素赋能人工智能创新发展[J].国家治理,2026,463(07):42-46.

[15] 潘小多,李新,冉有华,等.开放科学背景下的科学数据开放共享:国家青藏高原科学数据中心的实践[J].大数据,2022,8(1):113-120.

[16] 邹圣鹏,马福利,李云龙,等.面向空间科学卫星数据处理的资源调度算法[J/OL].计算机工程, 1-15[2026-08-13]. <https://doi.org/10.19678/j.issn.1000-3428.0260191>.

[17] 张思扬,胡林.融合标签语义的农业科学数据层次多标签分类方法研究[J/OL].数据与计算发展前沿, 1-10[2026-08-13]. <https://link.cnki.net/urlid/10.1649.TP.20260702.1803.004>.

[18] 孙苗,王子珂,童心,等.典型海洋环境观测数据产品应用现状及对我国的启示[J].大数据,2022,8(1):73-83.

[19] 沈志宏,张晓林,郑晓欢.PARIS原则:开放协作环境下科学数据的可用性[J].大数据,2023,9 (4):172-188.

[20] 骈澍晨,刘亚斌.甘肃省行业高质量数据集建设研究[J].甘肃科技纵横,2025, 54(07):51-60.

[21] 胡良霖,马宗文,朱艳华,等.我国科学数据发展回顾与思考[J].数据与计算发展前沿(中英文),2025,7(3):3-14.

[22] 朱艳华,高瑜蔚,胡良霖,等.我国科学数据标准规范实践与思考[J].中国科学数据(中英文网络版),2023,8(1):165-174.

[23] 胡良霖,朱艳华,李坤,等.科学数据伦理关键问题研究[J].中国科技资源导刊,2022,54(1):11-20.

[24] 朱艳华,廖方宇,胡良霖,等.科学数据安全标准规范关键问题探索[J].信息安全,2021,21(11):1-8.

[25] 陈祖刚,杨琳涵,何聪辉,等.人工智能就绪科学数据集分级与生产方法研究[J/OL].科学通报,1-13[2026-08-13].<https://link.cnki.net/urlid/11.1784.N.20260725.2020.002>.

[26] 高瑜蔚,胡良霖,朱艳华,等.国家基础学科公共科学数据中心建设与发展实践[J].科学通报,2024,69(24):3578-3588.

作者简介



高瑜蔚（1989-），女，博士，中国科学院计算机网络信息中心，高级工程师，主要研究方向为科学大数据要素化治理、标准规范与技术应用。

录用日期: XXXX-XX-XX

通信作者:

基金项目: 国家重点研发计划项目“基础研究科学数据国际联合服务网络建设合作研究”(No. 2025YFE0211200)

Foundation Items: National Key R&D Program of China “Research and demonstration on key technologies of collaborative services of multinational scientific data”(No.2025YFE0211200)