



文献 CSTR:

32001.14.11-6035.csd.2023.0044.zh

文献 DOI:

10.11922/11-6035.csd.2023.0044.zh

文献分类: 信息科学

收稿日期: 2022-12-02

开放同评: 2023-02-02

录用日期: 2023-03-22

发表日期: 2023-03-31

中国科学院科学数据中心体系建设实践及展望

陈昕^{1*}, 郑晓欢², 潘博雅¹, 沈志宏¹, 周国春¹, 褚大伟²

1. 中国科学院计算机网络信息中心, 北京 100083

2. 中国科学院科学传播局, 北京 100864

摘要: 科学数据已成为创新发展的基础性、战略性资源。科学数据中心作为科学数据长期保存和开放共享的重要载体, 在支撑科技创新发展中的作用日益显著。中国科学院面向新型科研范式和国家创新战略等新需求, 基于 40 多年的科学数据库建设积累, 经过 3 年的积极探索, 截至 2021 年底, 已建成“1 家总中心+18 家学科中心+13 家所级中心”的中国科学院科学数据中心体系, 数据资源总量约 100 PB, 基本实现了科学数据中心体系的常态化运行与服务, 有力支撑了大数据时代的科技创新。未来, 将进一步优化完善科学数据中心体系布局, 持续推进科学数据资源汇聚, 打造科学数据品牌, 发展数据智能的创新环境, 支撑新范式下的融合科学创新。

关键词: 科学数据; 科学数据中心; 数据管理; 数据共享

1 建设背景

1.1 科学数据管理面临新的挑战

随着科学数据资源的爆炸式增长, 数据密集型科学发现成为科学研究新范式。科学数据既是科学研究的重要产出, 也是新一轮创新活动的重要驱动力, 现已经成为科学研究创新和社会经济发展的战略性、基础性资源。但随着开放科学浪潮的快速发展^[1], 我国首个科学数据领域全国性的政策文件——《科学数据管理办法》的逐步落实^[2], 科学数据管理与应用也面临着新的挑战。

(1) 开放科学获得广泛共识, 面向科研活动的科学数据汇聚与管理需求迫切

传统的科学数据来源主要以大科学装置、观测网络、高通量计算等为主, 而前期的科学数据管理的主要对象也是这些“大科学数据”。而随着开放科学的全球化发展, 以信息技术发展为基础的科研数据开放共享被认为是下一个十年科研活动最显著的特征^[3], 科研人员日常科研活动产生的科学数据成为了科研产出的成果之一, 也成为了开展新的科学研究的重要输入, 这些“小科学数据”也被纳入了科学数据管理的全流程。《科学数据管理办法》中也对财政资助的科技项目产生的数据、论文关联数据的汇交管理等提出了明确要求。然而, 前期建设的科学数据共享服务平台对科技项目数据、科技论文关联数据、科学实验数据等科学数据的汇交管理关注不足, 缺少开放、可信的科学数据汇聚渠道和相应的制度机制及服务能力。

* 论文通信作者

陈昕: chx@cnic.cn

（2）数据密集型科研范式需要高质量、高可用的科学数据资源与服务

数据密集型计算成为新的科研范式，科学数据也成为了科研创新的动力源泉^[4]。而如何更好地发挥科学数据的价值，从未知中探索科学发现，还面临 3 个方面的挑战：首先，就科学数据本身而言，需要更加丰富的、可复用的开放科学数据。以中科院的科学数据为例，学科分布仍不够全面，且高质量数据相对较少，需要对科学数据资源体系布局进一步完善和发展。其次，在科学数据可用性方面，随着深度学习等新技术的快速发展和广泛应用，对科学数据的机器可读、机器可理解、机器可操作的服务能力要求进一步提升，实现“数据不下线”，自动化的按需服务。此外，挖掘利用科学数据，开展数据驱动的科学发现，需要相应的方法、工具和平台及稳定服务的载体以支持创新应用。

（3）数据纳入国家战略，需要相适应的基础制度和组织模式

数据作为新型生产要素，是数字化、网络化、智能化的基础，数据基础制度建设事关国家发展和安全大局^[5]。充分发挥海量科学数据价值，同样需要在数据确权、数据安全等方面形成科学数据的基础制度体系。同时，科研项目周期性的资助模式造成科学数据管理无法常态化、规范化地持续开展，新形势下也需要新的科学数据工作组织模式，构建起科学数据的基础设施和基础服务，可靠支撑科学数据驱动科技创新与发展。

1.2 科学数据中心建设是推动数据驱动创新的重要手段

科学数据中心作为科学数据长期保存的重要载体和开放共享的重要平台，逐渐发展成为科学研究的“加速器”^[6]，日益得到各方的关注和重视。欧美等发达国家高度重视科学数据资源的长期积累和合理利用，长期投入科学数据中心建设。美国在 20 世纪由美国国家航空航天局（NASA）主导建立的地球科学领域系列国家级数据存档体系 DAACs，长期为地球领域科研人员提供科学数据服务，成为世界领先的科学数据中心^[7]。英国剑桥晶体数据中心从 1965 年起持续收集数据，成为物质科学领域全球权威数据中心^[8]。这些发达国家通过多层面立法和政策支持、持续的多元化经费资助、同高水平科研机构和国际计划协同发展，推动科学数据中心的高质量运行，有效促进本国和地区的科技创新发展。

此外，近年来，世界各国和地区也强化科学数据中心的体系化布局和协同服务能力建设。欧盟部署的欧洲开放科学云计划（EOSC）联合欧洲现有的分布式科学数据基础设施，采用联邦式的协同模型，从架构、数据、服务、接口、规则、治理 6 个方面，打造一个开放的研究环境，提供跨境、跨领域的科研数据存储、管理、分析与再利用服务^[9]。德国自 2018 年起，推出国家科研数据基础设施计划（National Research Data Infrastructure Programme, NFDI）^[10]，以学科联合体为资助对象，按照学科领域布局国家科学数据基础设施。2020 年、2021 年分两批在化学、地质、地理信息、植物等基础学科，以及文本语料资源、人工智能、数据科学等前沿领域已布局 19 家科学数据基础设施，并计划于 2023 年 3 月发布第三批资助名单，其中包括提供科学数据基础服务的 Base4NFDI，强化国家科学数据基础设施的体系化服务能力。《科学数据管理办法》中也明确提出统筹推进国家科学数据中心建设和发展，将科学数据中心作为促进科学数据开放共享的重要载体，并指出科研院所等法人单位是科学数据管理的责任主体^[2]。

1.3 中国科学院科学数据中心体系是科技基础能力建设的重要组成部分

中国科学院作为国家重要的战略科技力量，长期关注科学数据管理与开放共享工作，一直是我国科学数据工作的排头兵。自 20 世纪 70 年代末启动科学数据库建设工作以来，经过多个五年规划

的实施,积累了丰富的科学数据资源,初步形成支持基础科研活动与科技创新的科学数据云环境,科学数据开放共享程度日益提高^[1]。此外,中国科学院拥有 100 多家科研院所和完整的自然科学学科体系,在推动我国科学数据基础设施建设和跨学科融合创新等方面可发挥示范引领作用。

在科学研究范式深刻变革、开放科学形成全球共识、科学数据对于国家创新发展日益重要的当下,面向资源汇聚等方面的现实问题,中国科学院需要常态化、规范化的科学数据管理与共享服务平台以支持科学数据的汇聚、推动科学数据的开放共享、保障科学研究的数据供给和技术支持。为此,中国科学院启动科学数据中心体系建设,发展体系化服务的科学数据基础设施,进而推动新型科技基础设施的建设与发展。中国科学院科学数据中心体系涉及研究机构众多、学科覆盖广泛,其建设也为我国科学数据基础设施的建设发展提供了重要的示范参考。

2 整体建设思路

为解决科学数据资源汇聚、服务和组织管理方面的诸多挑战,聚合力发展科学数据基础设施和数据驱动创新能力,中国科学院于 2019 年初启动中国科学院科学数据中心体系建设,并将科学数据中心体系作为全院科学数据管理与开放共享保障机制的重要组成部分^[12]。经历了 2 年的建设培育期,2021 年起,已进入常态化运行阶段,建立了全院科学数据制度规范体系和工作队伍,形成了包括“1 个总中心+18 个学科中心+13 个所级中心”的科学数据管理与共享的专业机构体系^[13],基本实现了全院科学数据工作的体系化推进,科学数据资源结构不断完善,资源总量约 100 PB,科学数据基础支撑服务能力得到显著提升,科学数据服务重大战略与重大工程、驱动科技创新等方面成效凸显,国际影响力进一步提升。2019 年 6 月,财政部、科技部公布 20 个国家科学数据中心^[14],中国科学院承建了其中 11 个,在国家科学数据工作中发挥了重要的作用。

2.1 建设思路

中国科学院科学数据中心体系旨在构建全院一体化的科学数据生产、管理、共享和应用的生态系统,持续推进科学数据资源积累和开放共享,挖掘科学数据价值,提升数据支撑服务能力,保障数据安全,促进科技创新和经济社会发展。

科学数据中心体系建设是一项复杂而综合的系统工程,包括了与科学数据相关的一系列概念、政策、方法、目标、系统等,需要从多维度进行设计,并涵盖科学数据管理的全生命周期和科学研究的全流程,以保证体系设计架构清晰、要素全面。为此,借鉴 DoDAF2.0 的架构设计方法^[15]以及国家信息化体系的建设经验^[16],从政策、规范、资源、网络、技术、应用、组织运行等方面规划和推动中国科学院科学数据中心体系建设。如图 1 所示,科学数据从科学研究中来,经过规范化的管理与处理,通过协同的科学数据服务网络支撑科研创新。体系的建设由政策管控,由规范指导,通过组织管理机制保障运行。

在具体建设层面,如何实现全院科学数据工作体系化、规范化、常态化,是中国科学院科学数据中心体系建设的主要挑战。通过从政策规范、组织管理、数据资源、技术工具、应用服务等多层次、多维度开展体系化建设,制定科学数据管理和科学数据中心管理的规范和原则,将科学数据的汇聚、管理、服务同科研活动及科研管理相结合,探索中国科学院体系化、规范化、常态化推动科学数据中心体系建设的发展路径。

(1) 多维推动体系化建设

政策规范的体系化。鉴于中国科学院实行院所两级管理，在贯彻国家相关政策基础上，将全院统筹规划和研究所责任制落实相结合，发展院所两级相协同的科学数据管理政策体系；针对全院学科覆盖广泛、领域科学数据管理具有差异性的实际情况，发展通用规范和领域规范相结合的规范体系。

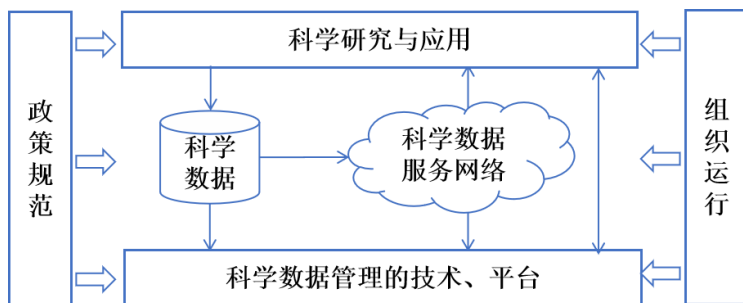


图1 科学数据中心体系架构设计思路

Figure1 Architecture design of scientific data center system

组织管理的体系化。在政策规范指导下，统筹科学数据中心布局，协调科学数据中心体系内的各类主体的目标和业务流程；发展科学数据相关的专家队伍和工作队伍，充分整合支撑科学数据中心建设的各类资源。

数据资源的体系化。推进科学数据的持续积累和加工整合，不断完善学科领域布局，发展学科领域数据资源体系，推动领域数据资源的系统化汇聚；构建互联互通的科学资源网络，强化互操作能力建设，实现全院科学数据资源的有序、动态流转。

技术工具的体系化。加强共性技术、基础软件栈和公共服务平台研发，并在基础软件栈基础上，发展领域特色、与科学研究密切相关的技术工具；整合全院资源，构建集约化发展、协同化服务的科学数据基础设施，强化科学数据中心体系的技术支撑能力。

应用服务的体系化。充分发挥中国科学院多学科优势，发展多学科融合的知识服务体系，推动交叉学科的融合创新；以数据为纽带，连通数据管理和科研创新，协同推进数据驱动创新发展。

(2) 规范化管理

科学数据规范化管理方面，在遵循 FAIR (Findable、Accessible、Interoperable、Reusable) 原则基础上^[17]，强化科学数据分类分级保护和规范化共享。

分等级：对数据进行必要的分级分类，明确各级别、各类别数据的管理要求和开放共享条件。

可发现：每个数据集都应被赋予符合国家标准的、持久化的唯一标识，并配有规范的元数据描述，易于发现和定位。

可访问：在保障数据安全的同时，开放共享的数据应公开提供稳定且易于获取的访问地址和访问方式。

可互操作：采用通用的信息记录格式和规范化词表，实现数据集信息的语法、语义互操作。

可重用：应明确数据使用的条件和要求，具有对数据生成及处理过程、数据质量等的详细描述信息，符合相关的标准规范，以便可以在不同应用中重复使用或和其他数据融合后使用。

在科学数据规范化管理基础上，研究构建科学数据中心能力成熟度模型，发展科学数据中心的认定指标体系和评价指标体系，采用认定和评价相结合的管理模式，推动科学数据中心的规范化管理。

(3) 常态化服务

落实法人单位对科学数据工作的主体责任。法人单位是日常科研工作的管理主体,充分发挥各单位对科学数据管理和开放共享的主体责任,依托法人单位建立本单位的科学数据管理机构,通过稳定运行的平台支撑科学数据工作的常态化。

推动常态化的科学数据汇聚和共享。持续稳定的资源汇聚是保持科学数据中心活力和长期服务能力的重要基础。在传统的由项目资助开展资源建设的基础上,明确并落实科技项目数据和论文关联数据的汇交与管理,以科学数据中心为依托,推动全院科学数据开放共享工作的常态化发展。

科学数据管理融入科研活动全流程,发掘科学数据管理的内驱动力。科研人员开展科学数据管理和开放共享的动因往往来自政策制度或资助方要求^[18]，“自上而下”的单向推动难以形成可常态化运转的生态体系。因此,中国科学院科学数据中心体系特别关注科学数据管理对科研创新活动的服务和支撑保障,提升科学数据质量和管理服务能力,使科学数据中心成为科研活动产生的科学数据的管理载体的同时,也成为中国科学院科研创新活动的重要助推器。

2.2 基本组成

国际上科学数据中心的类型通常以学科型和机构型为主,例如德国 NFDI 主要以学科进行布局,而美国的国家级数据存档体系 DAACs 以及高校的科学数据中心等主要以研究机构为单元开展科学数据中心建设。中国科学院科学数据中心体系建设同时考虑了科研院所的主体责任以及学科资源的聚集优势,并兼顾了公共基础性服务建设的必要性,设立了总中心、学科中心、所级中心,三类科学数据中心共同协作,成为中国科学院科学数据管理与开放共享保障机制的重要组成部分^[19]。

总中心负责为科学数据中心体系提供通用技术和共性标准规范支持,建设运行全院科学数据公共服务平台,推动科学数据交叉融合应用等工作,具体包括统筹协调与评价、通用支撑、公共服务与交叉应用。

学科中心负责学科领域数据汇聚、整编、挖掘分析、共享服务以及推动数据驱动科研创新等工作,保障科学数据安全,提升国际影响力。

所级中心负责本法人单位科学数据汇交、整编、质量控制、共享与数据安全等工作。

如图 2 所示,总中心负责通用技术支撑和公共服务平台建设,为各科学数据中心提供全面支持,同时支持院相关主管部门开展规划和管理。学科中心加强本学科领域资源的汇聚,重点发展优质数据产品和应用,支撑本领域科技创新;其中尚未成为国家科学数据中心的学科中心应进一步加强资源汇聚和服务创新能力,努力为国家科学数据中心建设贡献力量。所级中心主要负责本单位数据的汇聚、管理和共享服务,强调其组织管理和资源集成制度的完善;在所级中心建设过程中,将科研活动和科学数据管理有机融合,使所级中心和研究单元之间互为依托,共同发展;资源富集、能力较为突出的所级中心可参与学科中心的建设或向学科中心发展。

目前,中国科学院科学数据中心体系包括 1 个总中心,18 个学科中心和 13 个所级中心^[13]。2019 年在高能物理、基因组、微生物、空间、天文、对地观测、青藏高原、生态、冰川冻土沙漠、地球系统、基础学科公共等领域部署了第一批学科中心,这 11 家学科中心同时也承担了国家科学数据中心的建设任务;2020 年,面向国家战略发展需求,结合中国科学院优势资源,遴选了植物、海洋、凝聚态物质、化学化工、地球大数据、干细胞与再生医学、脑科学等 7 家院学科中心,进一步发展领域数据资源汇聚和服务能力。在所级中心建设方面,2019 年和 2020 年分别择优遴选了 8 家科学数据中心作为优秀所级中心鼓励其开展能力建设,其中植物研究所科学数据中心和海洋研究所科学数据中心在领域资源汇聚和服务等方面成效显著,发展成了学科中心。目前 18 家学科中心所在地区以

及涉及的学科情况参见表 1。

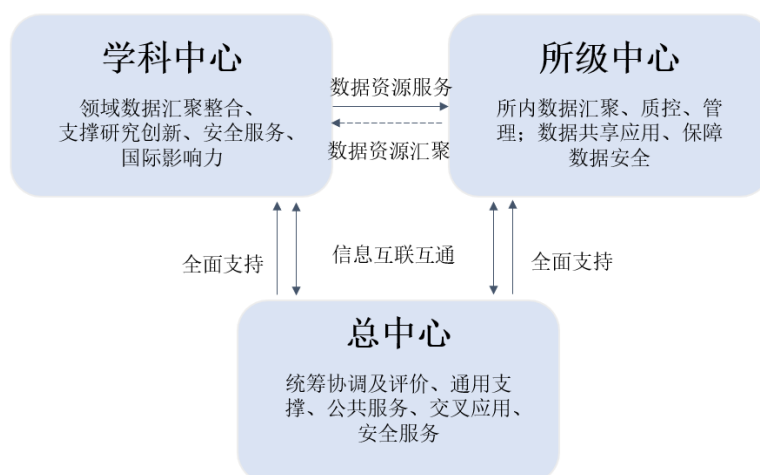


图 2 科学数据中心协同示意图

Figure 2 Diagram of scientific data centers collaboration

表 1 中国科学院科学数据中心体系学科中心名单

Table 1 Disciplinary centers list for the scientific data center system of Chinese Academy of Sciences

学科中心名称	地区	涉及学科 ^[20]
中国科学院高能物理科学数据中心	北京	高能物理学，天体物理学，空间天文学，天体演化学，星系与宇宙学，恒星与银河系，太阳与太阳系，天文学其他学科，粒子加速器，核科学技术其他学科，材料科学其他学科
中国科学院基因组科学数据中心	北京	生物学
中国科学院微生物科学数据中心	北京	微生物学，微生物分类学
中国科学院空间科学数据中心	北京	空间天文学，太阳与太阳学，空间物理学，标准科学技术，控制科学与技术，仿真科学技术，信息安全技术，电子技术，通信技术，人工智能，计算机应用，航空、航天科学技术基础学科，飞行器控制、导航技术，航空、航天材料，飞行器试验技术，航空航天地面设施、技术保障，航空、航天系统工程
中国科学院天文科学数据中心	北京	天文学
中国科学院对地观测科学数据中心	北京	雷达工程，摄影测量与遥感技术，地理学，森林经理学，大气科学，地质学，海洋科学
中国科学院青藏高原科学数据中心	北京	青藏高原地球系统科学研究领域，包括冰冻圈、水圈、陆地表层、大气圈、生物圈、固体地球、古环境、人地关系和遥感等
中国科学院生态科学数据中心	北京	大气科学，地球化学，地理学，水文学，海洋科学，生态学，植物学，昆虫学，动物学，微生物学，病毒学，人类学，农业基础学科，农艺学，土壤学，林业基础学科，环境科学技术基础学科，环境学，环境工程学，生态经济学，环境与生态统计学
中国科学院冰川冻土沙漠科学数据中心	兰州	地理学，地质学，水文学，地球科学其他学科，大地测量学，地图学，大气科学，地球化学，环境学
中国科学院地球系统科学数据中心	北京	大气科学，固体地球物理学，地理学，地质学，水文学，生态学，环境学，资源科学技术
中国科学院基础学科公共科学数据中心	北京	物理、化学、材料、动物、植物、病毒、核科学等基础学科和信息科学、干细胞代谢、脑成像等前沿研究领域，以及交通、智能电网、建材等行业领域

学科中心名称	地区	涉及学科 ^[20]
中国科学院植物科学数据中心	北京	植物学, 生物进化论, 生态学
中国科学院地球大数据科学数据中心	北京	地理学, 地球科学其他学科, 大气科学, 地图学, 海洋科学, 固体地球物理学, 摄影测量与遥感技术, 生态学, 动物学, 植物学
中国科学院海洋科学数据中心	青岛	海洋科学
中国科学院凝聚态物质科学数据中心	北京	凝聚态物理学
中国科学院脑科学数据中心	上海	神经生物学
中国科学院化学化工科学数据中心	北京	化学基础工程学科, 化学分类工程, 无机化学, 有机化学等
中国科学院干细胞与再生医学科学数据中心	北京	细胞生物学, 发育生物学

从地域分布来看, 32 家科学数据中心分布在北京、南京、兰州、昆明、广州、青岛、上海、长春、武汉、合肥等 10 个城市, 其中北京有 18 家 (含总中心和 15 家学科中心), 整体分布向科研基础设施和人才富集的区域集中, 同时也兼顾了区域的协同发展。

3 建设发展实践

经过 2 年的建设培育期, 2021 年起, 中科院科学数据中心体系进入常态化运行阶段。初步形成了院所两级的政策体系和规范化的科学数据中心管理制度, 科学数据资源不断发展, 建成了体系化的科学数据中心软件栈和公共服务平台, 在数据驱动创新和服务国家重大战略等方面发挥了重要作用。

3.1 政策制度与组织运行管理

政策、人才、运行机制是科学数据中心体系化发展的重要保障。中国科学院发挥了建制化优势, 形成了院所两级的政策体系, 并充分利用专家队伍, 保证科学数据中心体系的高质量统筹发展。

(1) 建立了全院科学数据政策规范体系

为规范和推动科学数据工作, 中国科学院及院属各部门、各单位制定并发布了相关的政策和制度 (表 2), 协同推进科学数据政策制度体系。

表 2 中国科学院部分科学数据相关政策

Table 2 Data policies released by Chinese Academy of Sciences

政策	发布时间	发布机构	规范对象
《中国科学院科学数据管理与开放共享办法 (试行)》 [12]	2019 年 2 月	中国科学院	科学数据工作
《中国科学院科学数据工作要点》 ^[21]	2020 年 12 月	中国科学院办公厅 (院网信办)	
《中国科学院科学数据中心管理暂行办法》 ^[18]	2021 年 9 月	中国科学院	科学数据中心 管理
《中国科学院科学传播局关于进一步加强和规范论文 关联数据管理的通知》	2021 年 9 月	中国科学院科学传播 局	论文关联数据 管理
《中国科协办公厅 中国科学院办公厅关于组织开展期 刊论文关联数据汇交工作的通知》 ^[22]	2022 年 9 月	中国科协办公厅 中国科学院办公厅	期刊论文关联 数据汇交

政策	发布时间	发布机构	规范对象
《中国科学院战略性先导科技专项管理办法》 ^[23]	2022 年 6 月	中国科学院	先导专项
《中国科学院战略性先导科技专项科学数据汇交管理实施细则（试行）》 ^[24]	2022 年 8 月	中国科学院	科技项目科学数据汇交

注：2022 年中国科学院网络安全和信息化领导小组办公室（简称“院网信办”）依托部门发生调整，由中国科学院办公厅调整为中国科学院科学传播局。

此外，科研活动数据管理和人才评价通常是由各研究所实际落实的，因此，院属单位管理制度被作为政策体系的主体。根据 2020 年度中国科学院信息化评估报告，截至 2020 年底，共有 63 家院属单位发布本单位科学数据管理制度。各单位结合自身情况，创新管理要求，切实将科学数据管理融入科研活动。东北地理与农业生态研究所将科学数据汇交纳入课题组绩效考核指标，水生生物研究所将科学数据汇聚和鉴定工作相结合，在推动本单位科学数据常态化管理方面发挥了重要作用。

（2）形成专家队伍和工作队伍

中国科学院科学数据中心体系建设充分发挥专家力量，形成常态化的工作机制，其管理架构如图 3 所示。其中，中国科学院网络安全和信息化领导小组是全院科学数据中心工作的领导和决策机构；院网信办是全院数据中心的日常管理机构；设立了科学数据中心专家委员会，为全院科学数据中心建设和运行管理工作提供咨询和指导；科学数据中心工作委员会作为科学数据中心建设和运行管理工作的支撑机构，主要由科学数据中心的专家组成，协助推动院数据中心体系建设发展的规划与实施，促进各中心之间的交流与合作；秘书处依托在总中心，作为科学数据中心体系日常办事机构。通过专家队伍和工作队伍的建立，推动体系管理工作的常态化规范开展。

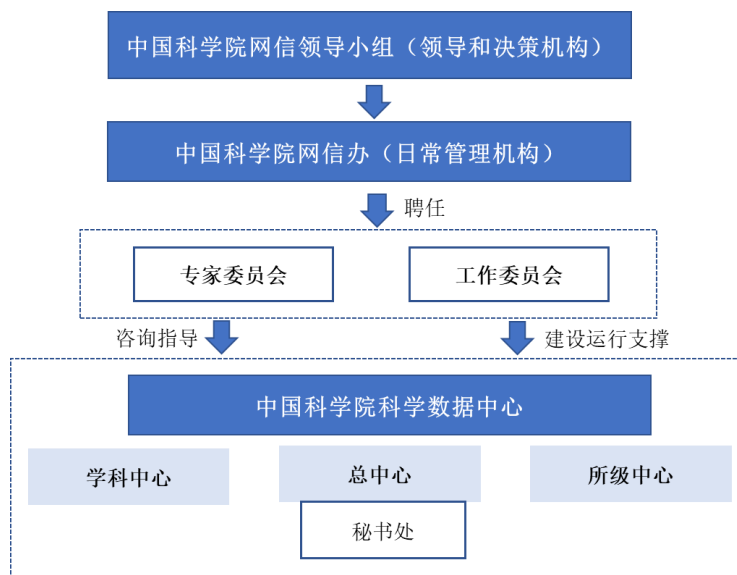


图 3 中国科学院科学数据中心管理架构

Figure 3 Management architecture for scientific data center of Chinese Academy of Sciences

（3）提出并实践了“建设-认定-运行-评价”的规范化全周期管理流程

《中国科学院科学数据中心管理暂行办法》提出了“建设-认定-运行-评价”的科学数据中心管理新机制。科学数据中心依据中国科学院科学数据中心建设指南及认定指标的要求建设，院网信办依据中国科学院科学数据中心认定指标，组织对认定申报单位开展认定评审，经认定通过的科学数

据中心，纳入中科院科学数据中心体系统一管理，常态化的运行服务，并参加中国科学院科学数据中心评价考核。评价考核结果为“不合格”的科学数据中心，限期整改仍不合格的，取消中国科学院数据中心资格。

为切实指导各研究所做好科学数据管理、开展科学数据中心建设，院网信办牵头编制了具有实操性的科学数据中心建设指南，规范科学数据中心建设，对科学数据中心组织管理、资源建设与管理、服务以及运行维护等方面的建设提供指南。并以科学数据中心建设指南为参考，结合中国科学院科学数据中心体系规范化管理要求，制定了《中国科学院科学数据中心认定指标》。2021年11月，组织开展了首批院科学数据中心认定。首批通过认定的中国科学院科学数据中心包括1个总中心，18个学科中心（其中11家为国家科学数据中心），以及13个所级中心。

在进入运行期后，为保障科学数据中心常态化运行并激励其积极发展，创新了多元的资源模式，将单一的建设经费资助模式调整为“运行维护（稳定支持）+应用示范类项目（竞争性争取）”相结合的方式对科学数据中心予以资助。在保证稳定运行的基础上，强化目标和需求导向。据统计，在2022立项的中国科学院网信示范项目中，有13家学科中心、11家所级中心从不同方向获得了资助。此外，通过常态化的运行监测和定期的评价考核，对科学数据中心建设进行管理和引导。对学科中心、所级中心进行分类评价，指标计量获取方式实现了定性定量相结合，力求评价结果客观且具有引导性，并将考核结果作为下一年度经费资助的依据。

3.2 互联互通的科学数据资源网络

较为完整的自然科学学科布局是中国科学院科学数据中心体系发展的重要优势，也是科学数据管理的巨大挑战。一方面，基于学科体系，可多渠道发展完善资源体系；另一方面，体系内资源的融合与互联互通也面临着互操作的难题。在建设过程中，通过学科中心遴选，扩展科学数据中心体系的学科架构；学科中心、所级中心分别从领域资源布局和研究所学科分布的视角规划其科学数据资源体系，并利用科技项目等渠道丰富了科学数据资源；总中心着重发展多层级的互操作规范，实现了科学数据中心体系资源和技术的互联互通。

（1）科学数据资源体系的不断完善

各科学数据中心，特别是学科中心梳理现有数据资源体系，不断拓宽数据资源覆盖的学科领域，注重发展前沿科学领域，加强综合服务能力建设，继续推动科学数据资源的持续积累、开放共享和综合利用。通过发展分中心等方式扩大资源汇聚，服务经济发展。特别是在支撑科技项目数据汇交方面持续发挥作用，仅2021年，中国科学院科学数据中心支持超过1000个项目开展数据汇交工作。截至2021年11月，18家学科数据中心数据总量达到94.40 PB，13家所级中心数据总量达到1.09 PB，其中学科中心是科学数据中心体系的资源主体。

从学科布局来看，18家学科中心主要按照学科领域进行布局，兼顾了区域和特殊环境。我们选取各学科中心所涉及的较主要的学科领域，统计不同领域所涉及的学科中心个数，并同国际数据存储库注册平台re3data上的存储库学科分布（截至2021年12月）做对比（如图4所示）。可以看出，中国科学院学科中心在地学领域占比最大，比re3data中的地学存储库占比高出约10个百分点，在物理、化学等基础研究领域具有较好的资源布局基础；但没有医学相关的学科中心，工程技术和前沿交叉类的学科中心也相对缺乏，领域的覆盖度还需要进一步完善。

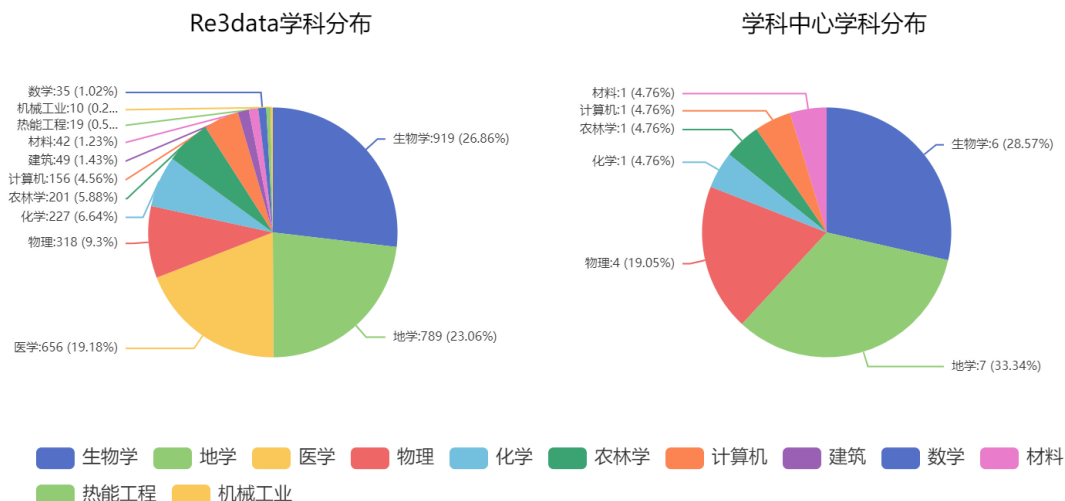


图 4 学科分布对比图

Figure 4 Comparative chart of discipline distribution

而随着科技项目数据汇交工作的推进，科学数据中心所汇聚的资源覆盖领域也在不断扩展，工程技术类资源发展迅速。图 5 是对各科学数据中心数据集所涉及的学科分类的统计（参照国标 GB13745-2009 学科分类与代码），一级学科后标注的数字代表涉及的科学数据中心的数量。可以看到，除图 4 中作为学科中心主要学科方向出现的地球科学、生物学、物理学、化学等一级学科外，工程与技术科学类的学科数据资源快速扩展，这也体现了当前前沿技术发展所关注的重点学科方向，将为科学数据中心体系资源发展带来新的生命力。

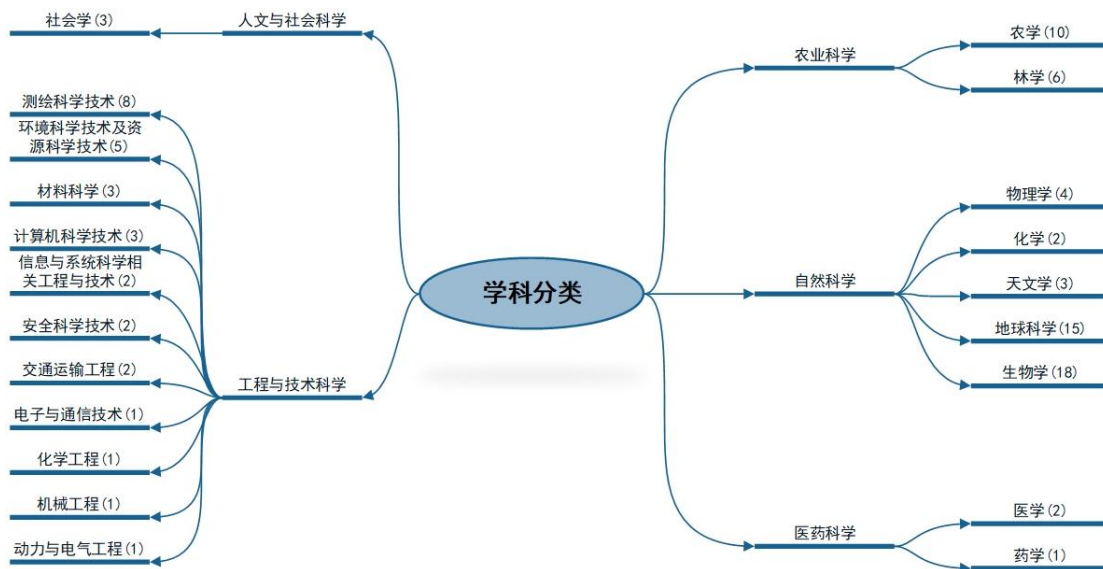


图 5 中国科学院科学数据中心体系数据集主要学科分布示意

Figure 5 Dataset distribution of major disciplines for the scientific data center system of Chinese Academy of Sciences

(2) 基于互操作规范的科学数据中心互联互通

中国科学院科学数据中心体系内各主体物理位置分散、学科领域多样、数据管理模式和数据规范存在差异，如何实现跨中心、跨学科数据资源的互联互通，实现数据资源的互操作，是体系化发展的重要挑战，而互操作性又恰恰是体现数据信息增值最大化的核心属性。为此，总中心从语法、语义、技术等方面实现科学数据中心体系内部数据资源的互操作。编制元数据术语规范，发布元数

据语义词表,提出元数据内容与汇聚要求,并通过互操作接口规范和收割认定系统来实现常态化、规范化的数据更新和互操作。

制定发布元数据术语与词表,支持跨领域数据的互操作性,提升科学数据中心体系资源开放性。面向中国科学院科学数据中心体系维护一组公共元数据术语,包括类、属性、词汇编码方案和语法编码方案,从而为中国科学院科学数据中心体系元数据标准、系统间的互操作提供基础,形成《中国科学院科学数据中心体系 元数据术语》^[25]规范。面向科学数据信息交换和协同计算场景下的科学数据以及科学数据中心等相关主体的元数据描述,发布符合 OWL 语法规则的 SD-INFOX 和 SD-COLLAB 词表,支撑和促进科学数据中心间的互操作。其中,SD-INFOX 词表包含数据集、人、机构、科学数据中心、科技项目等 5 个类和 80 个属性,通过标准模型和词表的使用,提高科学数据的可发现性以及跨机构联合查询的能力。SD-COLLAB 词表主要面向协同计算场景,对数据集、计算节点和数据表进行了规范。此外,词表的提出不仅面向中国科学院科学数据中心体系内的数据资源规范化管理和汇交,也同国际通用元数据词表 Schema.org、DCAT-V3 和 DCMI Terms 等词表建立了映射关系,对于推动中国科学院科学数据中心体系的开放性,有利于科学数据成果的影响力提升。基于术语和词表,发布《中国科学院科学数据中心体系 元数据内容与汇聚要求》^[26],明确科学数据中心间互操作的语法和语义表达。

同认证体系相结合的收割与验证,形成常态化的互通机制。建立了一套面向互联互通的科学数据中心互操作协议,已包括数据获取、数据发布相关接口 8 个,在保证全流程可追溯的前提下,实现数据中心系统间的数据流转、机器可操作。科学数据中心互操作协议在国际开放传输协议的基础上,符合中国科学院科学数据中心元数据规范要求,是科学数据体系化服务的重要基础,是开展协同服务的基础性规范,也是科学数据中心认定、科技项目数据汇交、目录报送等重要业务的基本要求。基于互操作接口,已实现各科学数据中心间数据集元数据、统计访问信息等的定时收割和互联互通。

3.3 体系化的科学数据中心软件栈和公共服务平台

科学数据中心建设过程中,有很多共性需求,为减少重复建设,简化各科研院所的工作,总中心研发了科学数据基础软件栈,支撑科学数据中心建设,加速科学数据中心的 FAIR 化;面向科学数据中心体系的全面流通和生态化发展,构建公共服务平台。软件栈和服务平台总体架构,如图 6 所示。通过构建科学数据中心基础软件栈和公共服务平台,提升科学数据中心规范化管理和体系化服务能力,实现全院科学数据资源的全面汇聚与跨领域融合,形成协同服务的科学数据资源网络,推动数据驱动融合科学的创新发展,进而促进科研范式变革升级。

研发科学数据中心基础软件栈,提升科学数据中心自身服务能力。面向科学数据中心全流程管理,实现其基础能力和流转能力的提升,充分发挥科学数据中心体系各个创新单元的服务能力。目前,已发布数据协同管理工具 DataSpace、机构数据存储库工具 InstDB 等 8 个工具(部分参见表 3),支持科学数据中心的数据汇聚、管理、发布、处理、分析、质控、安全等多个方面。其中,针对科学数据全生命周期流动的自动化与可溯源问题,自主研发的高吞吐科学大数据流水线系统 PiFlow,性能比国际主流系统 Apache NiFi 提升 2 - 7 倍^[27]。已发布的相关工具已在 20 多家研究所部署应用,有效减轻研究机构科学数据管理的信息化建设负担,提升了科学数据中心的数据管理、分析和安全保障能力。

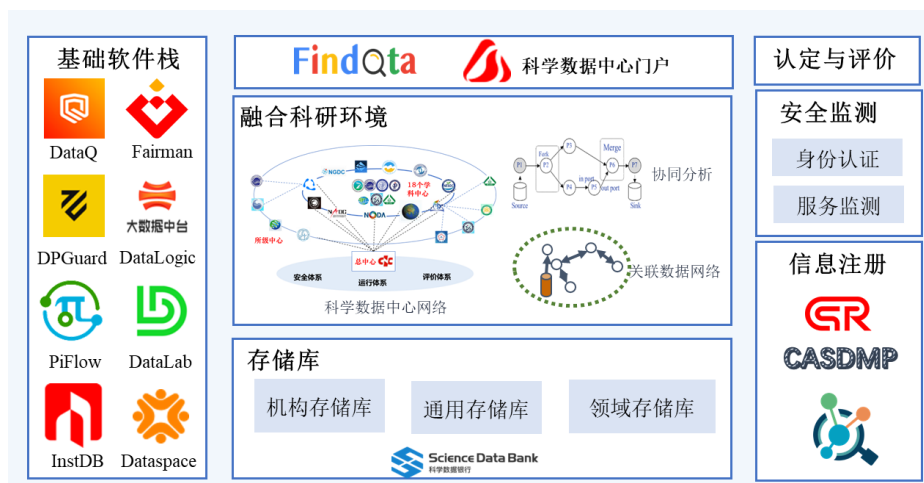


图 6 科学数据中心基础软件栈及公共服务平台

Figure 6 Basic software stack and public service platform for scientific data centers

表 3 科学数据中心软件栈列表

Table 3 Software stacks list for Scientific Data Center

工具	工具简介
机构数据协同管理工具 DataSpace	面向科研团队数据协作需求，为科研人员、科研团队提供安全、便捷的仓储式科研数据协同管理服务。科研人员可以在 DataSpace 创建空间，邀请成员进入空间，形成范围可控、多人协作的数据管理工作台，支撑团队内部的项目数据汇交、日常科研数据归档和管理工作。可以与机构存储库工具 InstDB 无缝集成，实现团队内部数据快捷发布。
机构数据存储库工具 InstDB	面向机构科研团队的数据公开发布需求，提供科学数据发布审核、资源发现、资源访问和用户支持等核心功能服务，通过构建科学数据中心标准化服务接口，实现多源数据的互操作，并支持数据资源唯一标识（DOI，CSTR）。
科学数据分析软件 DataLab	从底层算法开发调试到上层分析应用的一套软件框架。内置众多跨学科通用分析组件，具备组件分类管理和扩展能力，基于 Kubernetes 强大的资源调度能力为科研团队提供交互式分析和在线编程分析服务。
科学数据质量管理工具 DataQ	面向科学数据领域，提供数据校验服务，用于解决科学数据中心系统建设、运行及数据治理过程中的各种数据质量问题。
大数据流水线系统 PiFlow	一个基于分布式计算框架技术开发的大数据流水线处理与调度系统，将大数据采集、清洗、存储与分析进行抽象和组件化开发，以所见即所得、拖拽配置的方式实现大数据处理流程化配置、运行与智能监控。
数据门户安全防护软件 DPGuard	基于开源技术构建，为科学数据中心应用系统提供防注入、防篡改、防泄漏等安全防护，拥有多维度统计分析能力。

存储库是科学数据管理的通用基础性能力，总中心自主研发的科学数据银行 ScienceDB 现已获得施普林格自然（Springer Nature）、细胞出版社（Cell Press）、爱思唯尔（Elsevier）、泰勒-弗朗西斯出版集团（Taylor & Francis Group）等国际顶尖学术出版商推荐，可支持全球科研人员出版论文关联数据。各科学数据中心同 ScienceDB 合作，采用数据社区的发展模式，集约化建设论文关联数

据汇交的平台，推动数据开放共享社区生态建设的同时，也有助于形成品牌效应。目前，已建成数据社区近 200 个。

科学数据的流通是科学数据支撑创新的重要基础，为打通全院科学数据创新生态各环节，面向科学数据全局流通，以标识、科学数据目录、项目数据汇交、科学数据搜索、科学数据中心安全监测等服务为抓手，开展了面向科学数据体系的公共服务平台建设。其中，通过科技资源标识服务平台（cstr.cn）支持了全院科学数据资源的全局统一发现，该平台已同 DOI、Handle 等国际标识体系实现兼容；中国科学院项目科学数据汇交管理平台为科技项目数据汇交提供了统一的入口，全院科学数据汇交的全过程都可信息化的统筹管理；科学数据中心安全监控平台，每月为各数据中心提供安全扫描服务，为科学数据中心安全管理提供常态化的防护和保障。

面向科学数据中心间资源流通和融合的迫切需求，构建跨领域科学数据资源网络，打破“语义孤岛”，实现数据价值的深度挖掘和全面流通。面向重大科学问题和复杂应用场景，构建科学数据流协同服务平台，实现数据、计算、模型的跨中心调度与整合，在协同创新技术和模式等方面实现突破。

3.4 数据驱动创新的应用探索

中国科学院科学数据中心体系充分发挥自身资源优势与建制化优势，在重大任务保障方面合力攻关，同科研团队密切合作，强化科研活动全流程的科学数据管理，重点探索数据驱动创新的实施路径，在服务科技创新和支撑国家重大战略等方面取得了积极的成效。

目前，中国科学院已建成全球最大的天体光谱数据库 LAMOST，建成的全球微生物资源目录已成为全球微生物资源的数据仓储，吸引了来自 52 个国家和地区的 148 个保藏机构加入。海洋科学数据中心研制的全球海洋环境变化数据产品 3 年来被 28 份国际权威报告引用。地球大数据科学数据中心支持金砖国家可持续发展国际论坛数据发布，并为联合国可持续发展目标提供数据产品。

科学数据中心间紧密协作，支撑重大任务。高能物理科学数据中心和空间科学数据中心联合发布 GECAM 卫星观测数据，助力发现伴随快速射线暴的 X 射线暴。微生物科学数据中心面向新冠肺炎疫情防控 and 科研攻关，快速发布全球首个新冠病毒毒株信息及高清电镜照片，并建成了全球冠状病毒组学数据共享与分析系统，实现病毒组学数据集成与标准化的分析挖掘，总中心为其网络保障和数据分析可视化提供支撑。

在科学数据资源建设的基础上，发展特色数据产品和工具，并通过科学数据平台提供多元服务。地球系统科学数据中心构建了具有自主知识产权的集“数据、模型、知识”于一体的地理资源与生态知识服务平台，为我国重大科技工程、国家战略决策等提供了持续的地理资源与生态知识服务。冰川冻土沙漠科学数据中心构建的科学数据灾害应急响应服务平台已在 9 次灾害事件中提供科学数据应急响应服务。

中国科学院科学数据中心也在数据驱动创新方面积极探索。干细胞与再生医学科学数据中心团队围绕“衰老机制和调控”这一重要科学问题，汇聚整合多模型、多组学再生医学相关数据，建成跨越生命活动全周期的多维生物组学图谱数据，并基于数据成果分析揭示干细胞衰老、组织再生能力减弱与器官退行性变化的调控机制。南京地质古生物研究所科学数据中心构建的全球首个化石直翅目形态特征数据库揭示了中生代声学景观演化。此外，通过部署“大数据+AI”的范式变革示范项目，支持了生物、材料、大气、空间等领域数据驱动创新的研究探索，力求形成学科交叉的融合创新新路径。

4 发展展望

当前，国家创新发展步入新征程，中国科学院科学数据中心体系也将面临新的机遇和挑战。就发展现状而言，体系初步建成，但学科布局和国家战略及前沿领域间仍存在不匹配的情况，无法全面支持领域数据需求；且科学数据资源的可用性和科学数据平台的整体影响力仍显不足。随着 ChatGPT 等智能化技术应用的出现，数据智能相融合的创新驱动模式将是科学数据中心发展的重要关注方向。此外，在当前国际形势下，如何统筹安全与发展，在保留科学研究交流属性的基础上，开展科学数据的可信共享，也面临着机制、技术等多方面的挑战。为此，面向新型科研基础设施建设和高质量创新发展，未来需要在科学数据体系优化、高质量资源汇聚与发展、数据驱动创新发展以及数据治理和可信流通等方面进一步推动和发展。

4.1 优化科学数据中心体系，完善学科布局，鼓励多元化发展

面向新时期国家创新发展需要，聚焦基础研究和国际科技前沿，遵循学科领域特点和科学研究发展规律，完善科学数据中心的学科布局，形成学科体系更为完整、服务创新链条更加完善的中国科学院科学数据中心体系。在科学数据中心体系建设成果基础上，综合考虑科学数据库和科学数据平台与体系间的协同关系，实现以科学数据中心体系为核心、各类资源联动。

在发展路径方面，科学数据中心有着各自的优势和领域特色，可寻找适合自身特点的发展路径，发展学科型、区域型、应用型、科研型等多类型的科学数据中心。其中科研能力较强的科学数据中心，可以资源优势为平台，提升科学数据中心的 service 创新能力和自主创新能力，力争从资源汇聚中心发展成为科研创新中心。

4.2 进一步促进数据汇聚，提升数据质量，打造数据品牌

在数据汇聚方面，进一步全面落实和推进全院科技项目数据汇交和论文关联数据汇聚，延长汇聚服务链条，强化汇聚资源的管理、共享和引用追踪服务水平，切实保护数据作者利益，提升数据成果影响力，实现数据汇聚和利用的良性发展。同时，融入全球数据治理，积极参与国际开放科学政策的制定，主导和参与国际计划等国际合作，开展科学数据的国际交流与合作。

在数据质量提升方面，面向数据驱动创新和智能化发展目标，提升面向分析、应用等需求场景的科学数据可用性，发展数字化标准规范，加强机器可读、可操作、可理解的数据产品加工和数据服务建设，形成高质量数据服务能力。

在提升科学数据影响力方面，培育和打造若干具有国际影响力的科学数据中心；建设一批具有国际权威、具有重大战略意义或具有独具特色的科学数据库；进一步提升科学数据存储库的服务能力和国际影响力，发展成为国际科学数据资源的汇聚高地。

4.3 数据智能支撑融合创新发展

同科学研究单元协同发展。科学数据中心必须服务于科学研究，应同以全国重点实验室等国家科技力量相协同，发挥数据管理、数据分析、数据应用等方面的理论和技术优势，基于中国科学院科学数据中心网络支持国家战略科技力量构建。

提升数据支持融合创新能力。打造跨主体协同的数据计算一体化环境，发展高质量数据产品、分析模型及工具，实现数据、模型资源的有序、按需流转，形成以科学数据中心网络为依托的，资

源按需流转、可信计算的协同大科学实验室。

“大数据+AI”的范式变革，从示范到实践。在中国科学院网信专项应用示范“大数据+AI”范式变革类项目探索的基础上，总结摸索新范式下数据和智能融合支撑科技创新的理论方法和实施路径，并逐步推广，力争引领。发挥科学数据中心在新的科研范式下的数据纽带作用和支撑创新作用，形成中国科学院数据驱动科研发现的创新生态。

4.4 完善科学数据治理体系，强化科学数据安全

探索科学数据确权和价值分配机制。科学数据的权属和价值分配尚未明晰，一定程度影响了数据的开放和利用。利用科技资源标识、区块链等技术在数据确权、引用追溯、防篡改等方面的能力，探索中国科学院科学数据治理的新方法。

进一步完善科学数据中心评价机制。在现有科学数据中心能力认定和服务成效评价的基础上，进一步提升定量指标计量的获取能力，提升评价体系的规范性和影响力，参与国际认证体系合作。

强化科学数据安全。建立和完善科学数据全生命周期的安全管理制度和基础服务，将科学数据安全工作和科学数据管理工作深度融合，综合考虑狭义数据安全和广义数据安全，统筹安全与发展，在安全可信的前提下，分类分级开展数据要素化流通和应用服务。

作者分工职责

陈昕（1982—），女，博士，高级工程师，主要研究方向为科学大数据管理与处理分析。本文承担工作：数据分析与论文写作。

郑晓欢（1981—），女，硕士，高级工程师，主要研究方向为科学数据管理、科研信息化战略等。本文承担工作：论文研究思路、分析方法指导。

潘博雅（1988—），女，硕士，工程师，主要研究方向为科学数据管理与共享。本文承担工作：文献调研与数据分析。

沈志宏（1977—），男，博士，正高级工程师，主要研究方向为大数据管理与处理。本文承担工作：论文研究思路、分析方法指导。

周园春（1975—），男，博士，研究员，主要研究方向为数据挖掘与知识图谱。本文承担工作：论文研究思路、分析方法指导。

褚大伟（1979—），男，主要从事中国科学院网信发展规划、网信专项管理、科研信息化战略研究、数字中科院建设等工作。本文承担工作：论文研究思路指导。

致 谢

感谢中国科学院“十四五”网信专项“科学大数据工程（三期）”（CAS-WX2022GC-02）、“中国科学院科学数据总中心运行维护”（CAS-WX2022SDC-ZZX）项目资助。

参考文献

[1] UNESCO. UNESCO Recommendation on Open Science[R/OL]. (2021) [2022-11-22]. <https://en.unesco.org/science-sustainable-future/open-science/recommendation>.

- [2] 国务院办公厅关于印发科学数据管理办法的通知[EB/OL]. (2018-04-02). [2022-11-22]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm. [Notice of General Office of the State Council on publishing the Measures for the Management of Scientific Data[EB/OL]. (2018-04-02). [2022-11-22]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.]
- [3] MULLIGAN A, HERBERT R. Report: Research futures: drivers and scenarios for the next decade[R/OL].[2019-02-14]. <https://www.elsevier.com/connect/elsevier-research-futures-report>.
- [4] 郭华东. 科学大数据——国家大数据战略的基石[J]. 中国科学院院刊, 2018, 33(8): 768–773. DOI: 10.16418/j.issn.1000-3045.2018.08.001. [GUO H D. Scientific big data—a footstone of national strategy for big data[J]. Bulletin of Chinese Academy of Sciences, 2018, 33(8): 768–773. DOI: 10.16418/j.issn.1000-3045.2018.08.001.]
- [5] 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[EB/OL].(2022-12-19). [2022-12-19]. http://www.gov.cn/zhengce/2022-12/19/content_5732695.htm. [Opinions of the State Council on Data Elements Playing a More Important Role in Building a Data Infrastructure System[EB/OL].(2022-12-19) [2022-12-19]. http://www.gov.cn/zhengce/2022-12/19/content_5732695.htm.]
- [6] 汪洋, 周园春, 王彦桐, 等. 适度超前推动科研基础平台建设支撑我国高水平科技自立自强[J]. 中国科学院院刊, 2022, 37(5): 652–660. DOI: 10.16418/j.issn.1000-3045.20220408002. [WANG Y, ZHOU Y C, WANG Y G, et al. Promoting infrastructure construction in advance to support sci-tech self-reliance and self-strengthening at higher level[J]. Bulletin of Chinese Academy of Sciences, 2022, 37(5): 652–660. DOI: 10.16418/j.issn.1000-3045.20220408002.]
- [7] 高孟绪, 王瑞丹, 王超, 等. 关于国家科学数据中心建设与发展的思考[J]. 农业大数据学报, 2019, 1(3): 21–27. DOI: 10.19788/j.issn.2096-6369.190302. [GAO M X, WANG R D, WANG C, et al. Thoughts on the construction and development of national scientific data centers in China[J]. Journal of Agricultural Big Data, 2019, 1(3): 21 – 27. DOI: 10.19788/j.issn.2096-6369.190302.]
- [8] The Cambridge Crystallographic Data Centre (CCDC) [EB/OL]. [2022-11-22]. <https://ccdc.cam.ac.uk/>.
- [9] European Open Science Cloud (EOSC) Strategic Implementation Plan[EB/OL]. [2019-7-24]. https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/european-open-science-cloud-eosc-strategic-implementation-plan_en.
- [10] NFDI. Nationale Forschungsdaten Infrastruktur [EB/OL]. [2023-3-12]. <https://www.nfdi.de/?lang=en>.
- [11] 黎建辉, 周园春, 胡良霖, 等. 中国科学院科学数据云建设与服务[J]. 大数据, 2016, 2(6): 3–13. DOI: 10.11959/j.issn.2096-0271.2016061. [LI J H, ZHOU Y C, HU L L, et al. Scientific data cloud construction and service of Chinese Academy of Sciences[J]. Big Data Research, 2016, 2(6): 3 – 13. DOI: 10.11959/j.issn.2096-0271.2016061.]
- [12] 中国科学院科学数据管理与开放共享办法（试行）[EB/OL]. (2021-02-11). [2022-11-22]. <https://www.cas.cn/gzsd/ggsw/zh/202108/P020210827414998424076.pdf>. Measures for the Management of Scientific Data and Open Sharing of Chinese Academy of Sciences (Trial) [EB/OL]. (2021-02-11). [2022-11-22]. <https://www.cas.cn/gzsd/ggsw/zh/202108/P020210827414998424076.pdf>.
- [13] 中国科学院科学数据中心首批通过认定名单公布[EB/OL]. [2022-11-22]. <https://mp.weixin.qq.com/s/Jr51lTI-Nsa8J1hIE1-JDw>. [The first batch of approved list for Science Data

- Center of Chinese Academy of Sciences published [EB/OL]. [2022-11-22]. <https://mp.weixin.qq.com/s/Jr5l1TI-Nsa8J1hIE1-JDw.>
- [14] 科技部 财政部关于发布国家科技资源共享服务平台优化调整名单的通知[EB/OL]. [2022-11-22]. https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/qttw/qttw2019/201906/t20190610_147031.html. [Notice of the Ministry of Finance of the Ministry of S&T on Publishing the list of Optimizing and Adjusting the National S&T Resources Sharing Service Platform][EB/OL]. [2022-11-22]. https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/qttw/qttw2019/201906/t20190610_147031.html.]
- [15] The DoDAF Architecture Framework Version 2.02[EB/OL]. [2022-11-22]. <https://dodcio.defense.gov/library/dod-architecture-framework/>.
- [16] 郭诚忠. 中国信息化的发展思路 and 主要任务[J]. 中国科学院院刊, 1997, 12(6): 449–450. DOI: 10.16418/j.issn.1000-3045.1997.06.017. [GUO C Z. Development ideas and main tasks of informatization in China[J]. Bulletin of the Chinese Academy of Sciences, 1997, 12(6): 449–450. DOI: 10.16418/j.issn.1000-3045.1997.06.017.]
- [17] WILKINSON M D, DUMONTIER M, JSBRAND JAN AALBERSBERG I, et al. The FAIR Guiding Principles for scientific data management and stewardship[J]. Scientific Data, 2016, 3: 160018. DOI: 10.1038/sdata.2016.18.
- [18] Digital Science, Goodey G, Hahnel M, et al. The State of Open Data 2022[R/OL]. (2022-10-18) [2022-11-22]. <https://doi.org/10.6084/m9.figshare.21276984.v5>.
- [19] 中国科学院关于印发《中国科学院科学数据中心管理暂行办法》的通知[EB/OL]. [2022-11-22]. <http://www.ecas.cas.cn/gzwj/ygzdzd/202204/P020220429365585129260.pdf>. [Notice of the Chinese Academy of Sciences on publishing the Interim Measures for the Management of the Scientific data center of Chinese Academy of Sciences][EB/OL]. [2022-11-22]. <http://www.ecas.cas.cn/gzwj/ygzdzd/202204/P020220429365585129260.pdf>.]
- [20] 先导专项数据汇交实施在即, 院数据中心使命在肩. [EB/OL]. (2022-09-01) [2022-11-22]. https://mp.weixin.qq.com/s?__biz=MzkxODEzMDIwNA==&mid=2247507643&idx=1&sn=95e9b0ff565a359ad8702f5e734b8e5e&chksm=c1b485c8f6c30cde413cab9e4382de35232512e682c9606e4b340d18a36dded65fa54232f4f3&mpshare=1&scene=1&srcid=1226Po3v5pDQGqZdbdSca57O&sharer_sharetime=1672038550531&sharer_shareid=d4c8af816a4c2979df0b0765c54198fa&version=4.1.2.6017&platform=win#rd. [Scientific Data Centers CAS Are Ready For The Implementation Of The Strategic Priority Research Program. [EB/OL]. (2022-09-01) [2022-11-22]. https://mp.weixin.qq.com/s?__biz=MzkxODEzMDIwNA==&mid=2247507643&idx=1&sn=95e9b0ff565a359ad8702f5e734b8e5e&chksm=c1b485c8f6c30cde413cab9e4382de35232512e682c9606e4b340d18a36dded65fa54232f4f3&mpshare=1&scene=1&srcid=1226Po3v5pDQGqZdbdSca57O&sharer_sharetime=1672038550531&sharer_shareid=d4c8af816a4c2979df0b0765c54198fa&version=4.1.2.6017&platform=win#rd.]
- [21] 中国科学院办公厅关于印发《中国科学院科学数据工作要点》的通知[EB/OL]. [2022-11-22]. <http://www.digital.cas.cn/gzwj/ygzdzd/202101/P020210107504156285987.pdf>. [Notice of the general office of Chinese Academy Sciences on publishing Key Points of Scientific Data Work for Chinese Academy of Sciences][EB/OL]. [2022-11-22]. <http://www.digital.cas.cn/gzwj/ygzdzd/202101/P020210107504156285987.pdf>.]

- [22] 中国科协办公厅 中国科学院办公厅关于组织开展期刊论文关联数据汇交工作的通知[EB/OL]. [2022-11-22]. http://www.chuanbo.cas.cn/gzzd/202302/t20230202_4873798.htm. [Notice of the general office of China Association for Science and Technology the General Office of Chinese Academy of Sciences on Organizing the archiving of Scientific data related to Journal articles [EB/OL]. [2022-11-22]. http://www.chuanbo.cas.cn/gzzd/202302/t20230202_4873798.html.]
- [23] 中国科学院关于印发《中国科学院战略性先导科技专项管理办法》的通知[EB/OL]. [2022-11-22]. https://www.cas.cn/tz/202206/t20220608_4837579.shtml. [Notice of the Chinese Academy of Sciences on publishing Measures for Management of Strategic Priority Research Program of Chinese Academy of Sciences[EB/OL]. [2022-11-22]. https://www.cas.cn/tz/202206/t20220608_4837579.shtml.]
- [24] 中国科学院关于印发《战略科技先导专项管理实施细则》等 9 个文件的通知[EB/OL]. [2022-11-22]. http://www.iap.cas.cn/gb/jgsz/glxt/jhcwc/kyjfgl/zkyzd/202210/t20221014_6530477.htmlhttp://iap.cas.cn/gb/jgsz/glxt/jhcwc/kyjfgl/zkyzd/202210/t20221014_6530477.html. [Notice of the Chinese Academy of Sciences on publishing the Implementing Rules for Management of Strategic Priority Research Program of Chinese Academy of Sciences and other 8 files [EB/OL]. [2022-11-22]. http://www.iap.cas.cn/gb/jgsz/glxt/jhcwc/kyjfgl/zkyzd/202210/t20221014_6530477.html.]
- [25] 刘宁, 黄道然, 王栋栋, 等. 中国科学院科学数据中心体系 元数据术语[EB/OL]. 中国科学院计算机网络信息中心, 2021. <https://www.casdc.cn/md.html>. [LIU N, HUANG D R, WANG D D, et al. Metadata Terminology for Science Data Center System of Chinese Academy of Sciences[EB/OL]. Computer Network Information Center, Chinese Academy of Sciences, 2021, <https://www.casdc.cn/md.html>.]
- [26] 刘宁. 中国科学院科学数据中心体系元数据内容与汇聚要求(V2). 中国科学院科学数据总中心[EB/OL]. [2022-06-24]. <https://cstr.cn/39999.99.casdc.272CKTy7>. [LIU N. Metadata and Archiving Requirements of the Science Data Center System of the Chinese Academy of Sciences(V2). General Scientific Data Center of CAS[EB/OL]. [2022-06-24]. <https://cstr.cn/39999.99.casdc.272CKTy7>.]
- [27] 朱小杰, 赵子豪, 杜一. 模型驱动的大数据流水线框架 PiFlow[J]. 计算机应用, 2020, 40(6): 1638–1647. DOI: 10.11772/j.issn.1001-9081.2019101793. [ZHU X J, ZHAO Z H, DU Y. PiFlow: model driven big data pipeline framework[J]. Journal of Computer Applications, 2020, 40(6): 1638 – 1647. DOI: 10.11772/j.issn.1001-9081.2019101793.]

论文引用格式

陈昕, 郑晓欢, 潘博雅, 等. 中国科学院科学数据中心体系建设实践及展望[J/OL]. 中国科学数据, 2023, 8(1). (2023-03-31). DOI: 10.11922/11-6035.csd.2023.0044.zh.

Scientific data center system of Chinese Academy of Sciences: practices and prospects

CHEN Xin^{1*}, ZHENG Xiaohuan², PAN Boya¹, SHEN Zhihong¹,

ZHOU Yuanchun¹, CHU Dawei²

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

2. Bureau of Science Communication, Beijing 100864, P.R. China

*Email: chx@cnic.cn

Abstract: Scientific data have become a basic and strategic resource for innovation and development. As an important carrier of scientific data for long-term preservation and sharing, scientific data centers play an increasingly significant role in supporting the development of scientific and technological innovation. Based on more than 40 years of scientific database development and three years of active exploration, by the end of 2021, CAS had built a scientific data center system of "one general center, 18 disciplinary centers and 13 institutional centers". The total amount of data resources is about 100 PB, basically realizing the normal operation and service of the scientific data center system, and serving as a strong support for scientific and technological innovation in the era of big data. In the future, it will further optimize and improve the layout of the scientific data center system, continuously promote the archiving of scientific data resources, create a series of scientific data brands, develop an innovative environment of data intelligence, and support integrative scientific innovation under the new paradigm.

Keywords: scientific data; scientific data center; data management; data sharing