

生物医药研发数字基础设施体系建设研究

张建楠¹, 王晓杰², 沙雏淋³, 李莹莹¹, 李校堃^{4*}, 谭蔚泓^{3*}, 李兰娟^{5*}

(1. 浙江数字医疗卫生技术研究院, 杭州 311100; 2. 温州医科大学药学院, 浙江温州 325035; 3. 中国科学院杭州医学研究所, 杭州 310000; 4. 温州医科大学, 浙江温州 325035; 5. 浙江大学医学院附属第一医院传染病重症诊治全国重点实验室, 杭州 310003)

摘要: 生物医药产业正以前所未有的速度发展, 其研发数字基础设施体系的建设和完善是国家生物科技创新和健康中国建设的重要基石。本文论述了生物医药研发数字基础设施的内涵与新时期生物医药研发数字基础设施的分类和系统体系; 总结了美国、欧盟和我国生物医药研发数字基础设施战略布局的现状及发展路径, 梳理了我国生物医药研发数字基础设施建设在国际影响力和权威性、发展模式设计和效能、可持续建设、管理组织架构建设方面与欧美国家存在的差距及在顶层设计、管理治理方面面临的挑战。研究建议, 加强生物医药研发数字基础设施顶层设计, 加强生物医学和健康医疗资源的统筹编排, 加强单一个体平台集群化整合建设和治理, 加强开放平台能力建设和运营保障。

关键词: 生物医药; 研发数字基础设施; 综合集成平台; 数据平台; 计算平台

中图分类号: R1 **文献标识码:** A

Digital Infrastructure System for Biomedical Research and Development

Zhang Jiannan¹, Wang Xiaojie², Sha Chulin³, Li Yingying¹, Li Xiaokun^{4*},
Tan Weihong^{3*}, Li Lanjuan^{5*}

(1. Institute of Medical-care Information Technology, Hangzhou 311100, China; 2. School of Pharmaceutical Sciences, Wenzhou Medical University, Wenzhou 325035, Zhejiang, China; 3. Hangzhou Institute of Medicine, Chinese Academy of Sciences, Hangzhou 310000, China; 4. Wenzhou Medical University, Wenzhou 325035, Zhejiang, China; 5. State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China)

Abstract: The biomedical industry is developing at an unprecedented speed and the establishment and improvement of a digital infrastructure for biomedical research and development (R&D) is crucial for biotechnology innovation and Health China construction. This study analyzes the implication, classification, and systematic framework of the digital infrastructure for biomedical R&D in the

收稿日期: 2023-09-13; **修回日期:** 2023-10-15

通讯作者: *李兰娟, 浙江大学医学院附属第一医院传染病重症诊治全国重点实验室教授, 中国工程院院士, 主要从事人工肝、传染病、数字医疗等研究工作;

E-mail: ljl@zju.edu.cn

*李校堃, 温州医科大学教授, 中国工程院院士, 主要从事以生长因子为代表的蛋白质药物基础理论与新药研发研究;

E-mail: xiaokunli@163.net

*谭蔚泓, 中国科学院杭州医学研究所教授, 中国科学院院士, 主要从事生物化学、生化分析、功能核酸和分子医学等研究工作;

E-mail: tan@hnu.edu.cn

资助项目: 中国工程院咨询项目“生物医药产业人工智能大数据集成平台发展战略研究”(2022-DFZD-28)

本刊网址: www.engineering.org.cn/ch/journal/sscae

new era. It summarizes the strategic layouts and development paths of the digital infrastructures for biomedical R&D in the United States, the European Union, and China, and examines the gap between China and European and American countries in terms of international influence and authority, development model design and efficiency, sustainability, and management organization construction. Moreover, it presents the challenges faced by the digital infrastructure of China regarding top-level design, management, and governance. Furthermore, we suggest that the following aspects should be strengthened: top-level framework design, coordination of biomedical and healthcare resources, clustered construction and governance of the platform, and capacity building and operation guarantee of the open platform.

Keywords: biomedicine; digital infrastructure for research and development; integrated platform; data platform; computing platform

一、前言

21 世纪初, 开放科学运动在国际社会迅速兴起, 引发了学术界、政府和产业界对研究数字基础设施发展的新思考^[1,2]。中国、美国、英国、德国、法国、加拿大和澳大利亚等多个国家成立了国家级开放科学中心并推动建设国家级研究数字基础设施作为配套基础设施进行资源的整合、管理、服务和二次重用, 支持信息资源开放集成和多边合作^[3,4]。生命科学、生物科学、医药科学等是研发数字基础设施建设的重要领域, 开放科学运动为生物医药研发数字基础设施向结构性、系统性整合化和开放生态化发展奠定了重要基础。2021 年, 联合国教科文组织 (UNESCO) 第 41 届会议审议通过了《开放科学建议书》, 标志着开放科学迈入全球共识的新阶段。193 个国家就包括开放科学研究数字基础设施建设发展在内的开放科学核心支柱内容达成共识, 并提出投资、协调、统一开放科学基础设施和服务目标^[5]。随着开放科学的发展, 许多早期建成的或同期开发的数据库系统和信息资源平台在各自领域内被逐步整合至单一或是分布式的国家级集成平台体系, 并进行系统性管理。典型的有国家科技资源共享服务平台、国家人口健康信息平台、中国医药信息查询平台等。

近年来, 在“数据科学”战略、“人工智能”战略、“健康中国”战略和“数字经济”战略等多元战略背景下, 大数据、云存储、算力等各种资源呈现爆发式增长, 领域关联、资源关联、平台互通协同的生物医药研发数字基础设施的发展对生物医药产业科学生产力和经济发展产生了深远影响。以应用需求为导向的生物医药研发数字基础设施正在成为支撑区域生物医药产业创新研发的重要利器, 支持产业发展在国际竞争和合作中占据战略性领先地位。目前, 世界范围内覆盖生命科学基础研究和健康医疗临床研究全领域研发要素、资源并且具备

多边协调耦合能力的数字基础设施相对较少。由于人体健康隐私安全、知识产权等方面的限制, 生物医药研发基础设施系统发展仍处于起始阶段^[6]。因此, 认清生物医药研发数字基础设施在新时期的战略定位, 并围绕其开展建设、运营等相关研究和探索实践必要且迫切^[7]。

本文是“生物医药产业人工智能大数据集成平台发展战略研究”项目的研究成果, 分析了生物医药研发数字基础设施的内涵、分类和层级系统, 梳理了中国、欧盟、美国等典型国家和地区在生物医药研发数字基础设施方面的战略侧重点和发展现状, 分析了我国在生物医药数字基础设施发展方面的差距及面临的挑战, 并提出相关措施建议, 以期为新时期生物医药研发数字基础设施科学、系统建设和可持续运营提供支撑和参考。

二、生物医药研发数字基础设施的内涵、分类和层级系统

(一) 生物医药研发数字基础设施的内涵

《开放科学建议书》中对开放科学基础设施的定义为支持开放科学和满足不同体系需求所需的虚拟的或物理的共享研究基础设施。包括主要科学设备或成套仪器、知识型资源, 如汇编、期刊和开放获取出版平台、存储库、档案和科学数据、现有的研究信息系统、用于评估和分析科学领域的开放文献计量学和科学计量学系统、能够实现协作式和多学科数据分析的开放计算和数据处理服务基础设施以及数字基础设施。我国国家发展和改革委员会最新对创新基础设施的定义为: 支撑科学研究、技术开发、产品研制的具有公益属性的基础设施, 包括重大科技基础设施、科教基础设施、产业技术创新基础设施等。在我国, 生物医药研发数字基础设施可定位为一种数字化的创新基础设施。

新时期, 生物医药研发数字基础设施泛指用于

支持生物医药研发和科学发现的研究资源和研发能力集成平台设施。区别于传统专注于实现数据资源的集成、融合与管理^[8]的数据库系统，新型基础设施专注实现数据资源、计算资源等必要研发能力的集成、融合、管理与服务的资源和研究能力服务平台^[9]。通过融合应用人工智能、大数据、云计算、云存储、数字孪生等系列新兴数字化技术，整合重塑已有平台或是搭建新平台来强化数据资源、工具集资源及生物医药大型设备的软硬件集成应用能力。新型生物医药研发数字基础设施平台的基本组成包括了网络基础设施和基于云的计算基础设施、研究资源基础设施和应用服务基础设施。生物医药领域资源应涵盖生命科学、生物科学、医疗健康等在内的数据资源及相应的工具、标准、知识、软件和硬件设施。

(二) 生物医药研发数字基础设施的分类

生物医药研发过程与生命科学、健康医疗和生物医药等领域密切相关。按照整合规模和整合目的的不同，现有数字化平台基础设施大致可分三大类：一是资源整合平台。以数据资源、工具资源、计算资源、学习资源、互操作资源等研发要素为轴心整合不同类型资源，形成“大而全”的资源查找、获取和应用服务平台。二是特定领域学科和主题式研究基础设施^[10]。以学科或研究主题为轴心集成或整合研发要素资源，形成“专而精”的主题式研究平台。三是综合型集成基础设施。整合以上两种平台形成的呈伞状、多层次结构的基础设施平台，强调资源域、技术域、应用域和服务域的规模化集成和融合。生物医药研发数字基础设施作用的发挥及其影响力的广泛建立通常取决于工具资源、数据资源、云计算资源等可用性和广泛适用性；可互操作的资源和数据标准的深度应用；平台功能如程序、工具、数据资源等的多样性、互操作性和可拓展性；云计算、云存储、人工智能、大数据、区块链等多项新一代信息技术的融合应用程度^[11]。全球生物医药研发数字基础设施主要分类及案例见表1。

(三) 生物医药研发数字基础设施层级体系

仿照生态学理论和对生态系统个体、种群和群落的定义，以“集成平台”为最小单位，国际国内现有生物医药研发数字基础设施已呈现出3个层

级：① 生物医药研究型集成平台，相当于“个体”，是指单个的生物医药研究数字基础设施；② 生物医药研究型集成平台群，相当于“种群”，即在物理空间内存在的同一领域、同一主题或同一类型的生物医药研发数字基础设施虚拟集群；③ 生物医药研究型集成平台集群，相当于“群落”，它们由物理空间内众多类型不同、规模层次、领域不一的生物医药研究型集成平台群通过完全整合或必要连接形成一个庞大且复杂的平台系统虚拟集群（见图1）。

生物医药研究型集成平台是研究机构、医院、高校、企业等构建并维护特定领域学科或专题的基础设施平台，如呼吸性疾病研究中心以及地方政府或国家基于社会成本和效益考虑直接统一构建的公共资源基础设施平台（如罕见病数据资源平台）。主流平台多是基于多源、多模态的生物医学大数据资源平台，通过集成技术叠加数据挖掘、分析、处理、流通、应用等能力，同时融合了应用云和超算等算力平台能力。研发活动在云平台工作流中的无缝衔接，使研究人员跨实验室和机构访问各种数据库、知识库、其他资源和工具更为方便，高性能计算能力支持快速完成从实验室到跨组织研究所有阶段中的计算分析任务^[12]。利用云计算/超算等计算资源并依托云存储等完成所有数据资源的经济存储、流通和应用；通过基于云的集成平台即服务（iPaaS）能力集成生物医药研发所需的数据资源、算力资源、研发工具与服务。

生物医药研究型集成平台群一般规模较大，呈树状结构，拥有一定数量的子节点平台，亦或是呈去中心化网状平台群分布。由于投入和运营成本巨大，通常是由政府部门、权威研究机构、非营利组织等合作共同建设。常见的有两类：一类是以层级化方式构建，例如，我国的人口健康信息平台在物理空间以“国家一省一市”三级进行构建；另一类是以网络化方式构建，多是国家级或世界级的主题式研究型集成平台，例如，美国合成生物学工程化平台合成生物学网络、世界卫生组织（WHO）国际病原体监测网络等。

随着生命健康和医疗卫生领域相关性研究和学科交叉研究的不断深入，生物医药企业和研究机构对于研发数字化资源的需求不再仅满足于单一资源的获取，对于相关但独立资源的互连、组合和整合需求愈发突出^[13]。数据资源的交互开始频繁发生于

表 1 全球生物医药研发数字基础设施的主要分类及案例

类型	说明	具体分类	案例
资源整合平台	平台以资源为中心架构	数据资源	英国生物银行 (UK Biobank)、药物银行 (DrugBank)、美国国家生物技术信息中心 (GenBank)、癌症基因组图谱 (TCGA)、日本 DNA 数据库 (DDBJ)、蛋白质结构数据库 (PDB)、药物遗传学和药物基因组学数据库 (PharmGKB)、在线人类孟德尔遗传数据库 (OMIM) 等
		算力资源	欧洲开放科学云 (EOSC)、中国科技云 (CSTCloud)、超算平台、高性能计算平台等
		工具资源	集群操作系统 (Galaxy)、参考管理软件 (Mendely)、英伟达智能计算平台克拉拉 (NVIDIA Clara)、云数据平台 (Databricks) 等
		教学培训资源	综合学习平台 (Coursera)、中国生物医药教育网、医学教育平台 Medscape、MedlinePlus、MedEdPORTAL 等
		知识资源	医学图书馆检索系统 (PubMed)、化学模组数据库 (PubChem)、中国生物医学文献服务系统等
		标准资源	生物医学数据交换基础设施 (i2b2 tranSMART Foundation)、生物途径交换标准 (BioPAX)、临床数据交换标准协会 (CDISC) 开发的临床数据获取协调标准 (CDASH)、观察性健康数据科学和信息学协作组 (OHDSI) 开发的通用数据模型 (OMOP CDM)、人类表型本体 (HPO) 等
		其他资源	学术分享研究社区 ResearchGate、Academia.edu 等
特定领域学科主题式研究基础设施	平台以学科和研究领域为中心架构	学科研究	合成生物学 (Biofoundries)、神经影像工具和资源合作实验室 (NITRC)、神经信息学 (COINS)、创伤性脑损伤研究信息学 (FITBIR)、人类代谢组学 (HMDB)、基因组学等
		疾病研究	人类功能交互网络 (ConsensusPathDB-human)、帕金森研究网络 (PDBP)、阿尔兹海默研究网络 (GAAIN)、肺结节研究网络 (LIDC-IDRI)、自闭症研究网络 (NDA) 等
		药物研究	外泌体 (ExoCarta)、分子交互研究 (IntAct)、药物临床试验登记与信息公示平台等
综合型集成基础设施	主题研究平台域和资源平台域嵌合架构	综合型	Elixir、NCBI、Australian BioCommons、de.NBI、国家科技资源共享服务平台等

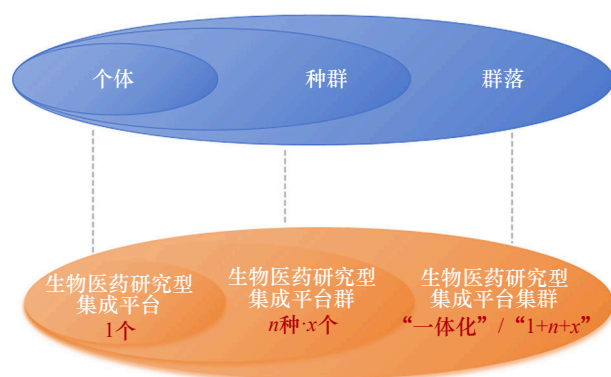


图 1 生物医药研发数字基础设施生态体系层级化示意图

两个或多个生物医药研究型集成平台群之间。生物医药研究型集成平台集群是超规模、多域共生的数

字综合体，支持各个研发活动的参与方能够在一个庞大的单一平台内快速、稳定地获取高质量的资源和研发能力，并快速实现数据向知识、产品和产业的价值转化。集群化、生态化发展是各国基础设施建设竞争的制高点。例如，欧盟的欧洲生命科学研究基础设施 (ELIXIR) 平台正在通过集成欧盟范围内的高质量平台力量，探索构建超规模化的基础设施集群生态。

三、欧美生物医药研发数字基础设施战略布局现状

美国和欧洲的生物医药产业最发达，创新研发

也最活跃，是生物医药卫生健康信息化建设的领先国家和地区。以美国、欧盟等为代表的国家和地区为保持其在世界范围内的技术和产业主权，积极布局建设新的生物医药研发数字基础设施以迎接生物医药产业和数字产业融合下新的产业机会。生物医药研发基础设施集群发展是多重因素引导下自然迭代进化的结果，美国和欧盟的模式呈现了两种典型的演化路径。

（一）美国

在美国，国立卫生研究院（NIH）是生物医药研发数字基础设施建设与发展的主要领导者。在NIH总体统筹下，生物医药研发数字基础设施战略布局呈现研发能力集中化、资源要素集中化、组织机构网络化等特征。

在基础设施技术方面，2017年12月，美国卫生与公众服务部（HHS）发布《生物医学发现与数据健康平台2017—2027年战略计划》^[14]。该战略计划提出要实现更大规模的生物医学数据可访问、互操作和可重用，加速将数据和信息转化为知识和洞察力。2018年9月，《数据科学战略计划》发布对数据基础设施、数据生态系统的现代化建设、数据管理、分析方法和工具的战略计划^[15]。NIH数据科学战略办公室2018年发布《发现、实验、可持续性科学和技术研究基础结构（STRIDES）计划》^[16]，提出通过使用大规模云计算平台（用于数据存储、访问和计算的共享环境）实现分布式数据存储资源的可访问性和规模经济。“美国国立卫生研究院战略计划（2021—2025年）”提出要资助一系列不同类型的研究项目来促进机器学习技术，并用于支撑大规模数据库的建设和管理^[17]。

在基础设施建设方面，美国生物医药研发数字基础设施的建立和使能通常与相关组织机构推出的大型研究计划的开展密切相关，作为项目配套的基础设施建设成果。例如，癌症登月计划、脑科学计划^[18]、人类基因组计划、人类生物分子图谱项目等均建设了相应的全球开放共享平台。近年来，美国又先后新增建设了11个生命大数据技术研发中心，通过增加对生物特征数据、生活方式和环境数据等非传统来源的新型数据的收集、研究和利用来促进药物开发。2022年，NIH成立高级健康研究计划局（ARPA-H），预算拨款65亿美元专注于解决从分子

到社会各个层面的问题，实现对癌症、传染病和阿尔兹海默病等多种疾病的预防、诊断和治疗，并为所有患者提供公平服务^[19]。与生物医药研发基础设施相衔接的是针对具体研发场景的小型集成设施单元。基于人工智能的敏捷研发模式在合成生物学领域开展探索。2016年，美国能源部生物能源技术办公室（BETO）成立了合成生物学和工程生物学敏捷生物铸造厂（Agile BioFoundry）^[20]，其通过提供一个集成的基础设施，应用人工智能/机器学习来增强迭代实验室的“设计-构建-测试-学习”生物工程循环和转向商业化生产的工艺研发创新。

在基础设施管理运营方面，生物医药研发数字基础设施的发展总体与HHS、NIH为主的相关机构、组织架构体系的设置协同。生物医药研发数字基础设施由NIH统筹管理运营，NLM则是实际开展数字资源管理和运营的核心机构，相关科学知识在成果发表或项目验收阶段将汇集到NLM。NLM下属的国家生物技术信息中心（NCBI）负责标准化收集、存储、管理和分发与生物技术相关的信息资源和工具资源，主题式分类汇集跨源的生物医药研发数字基础设施资源。目前，NLM已汇总包括NIH下设的27个研究所/中心实际运营的73个生物、医药、健康科研信息数据库，以及其他机构的300余个生物医药领域数据库及NIH非科研信息类数据库。NCBI支持和开展生物医学信息学和健康信息技术的研究、开发和培训，并负责协调由6500个成员组成的全国医学图书馆网络，为美国各社区提供医学健康信息，包括生物医药科学出版物、研究数据、元数据、开放式教育资源、软件以及源代码和硬件。作为支撑，NIH曾在2012年大数据到知识（BD2K）项目支持下开发了BD2K-LINCS DCIC平台，通过构建一个高容量、可扩展的集成知识环境，实现对所有LINCS资源以及来自其他相关资源的许多其他外部数据类型进行联合访问，实现直观查询、集成分析和可视化。

美国NCBI建立了具有国际影响力的生物医药数据资源平台，主导着国际生命科学的研究。NCBI和食品药品监督管理局（FDA）等组织机构从生物医药研发至监管准入虹吸了全世界包括中国在内的众多生物医药方面的宝贵科学知识、实验数据以及临床数据。

(二) 欧盟

欧盟围绕生物医药研发数字基础设施实施基于协调和连接的综合集成发展路径。2020年,《塑造欧洲的数字未来》^[21]《欧洲数据战略》和《欧洲人工智能白皮书》等战略文件和报告在基础设施和工具、标准和法律规则、价值观和社会模式三个层面展开人工智能和大数据技术战略部署。2021年1月,欧盟发布实施第九期研发框架计划——“欧洲地平线”计划(2021—2027年),预算投入24亿欧元投资建设ELIXIR平台,旨在汇集、管理和共享全欧洲的生物与医疗大数据、分析工具、知识成果等,在平台建设和运营、国际合作关系建立和维

护、最大化平台作用发挥等多个方面进行了科学的制度设计^[22]。

ELIXIR通过链接、集成和集群成熟的资源、研究社区和各类能力平台,向研究社区提供综合的数智化研发能力,如图2所示。由欧洲生物信息学研究所(EMBL-EBI)作为枢纽担任协调秘书处角色。ELIXIR数据平台共收录了19个核心数据资源库,包括高通量功能基因组实验数据、人类蛋白质编码基因信息、罕见疾病相关高质量数据集、基于质谱的蛋白质组学数据、蛋白质测序数据和其他资源;ELIXIR收录了星系(Galaxy)平台、欧洲开放科学云(EOSC)等高质量计算资源;ELIXIR标

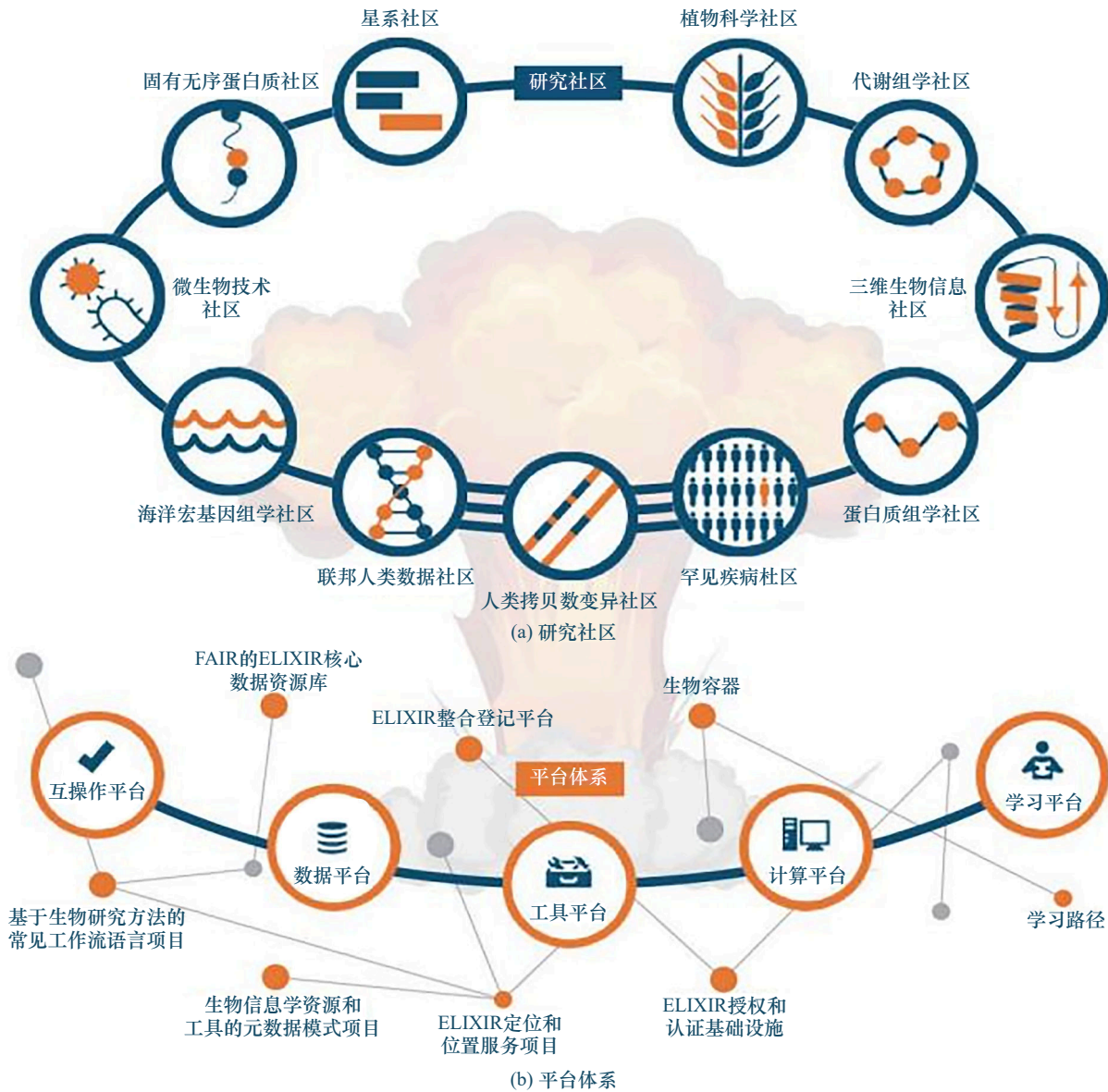


图2 欧盟ELIXIR基础设施示意图

准平台应用 Biotoools Schema 从来源、查询和发现实现标准化跨软件项目、存储库、计划和组织的元数据的管理和交换；开发简易文档相册和管理器 (EDAM) 本体^[9]支持对工具功能进行严格的一致性描述，帮助用户快速找到、理解和比较相关工具。使用 EOSC 构建基于数据和云服务的供应链产业的活力生态系统，促进成果共享，支持建立医药制造产业和医疗数据通用欧洲数据空间。基于 ELIXIR 的生物医药研发合作正在各个研究社区快速开展，其合作圈在全球范围内持续拓宽^[23]。

ELIXIR 的成功之处在于保障基础设施有效运作的各项机制建设完善。在研发合作方面，平台集群化后采取了网络公示等透明化手段，促进研发、转化、监管范畴内的资源配置快速高效进行，加速不同研究社区的研发合作。ELIXIR 平台与创新药物计划 (IMI)、欧盟战略研究议程 (SRA) 等共同发展，积极发挥基础设施作用。在首期《“欧洲地平线”计划 2021—2024 年战略计划》^[24]中，优先发展的领域包括：化学品风险评估、健康欧洲研究区、卫生保健系统转型、个性化医药、罕见病、抗菌素耐药性。SRA 列出了开发和整合下一代数字和生物医学技术的优先事项——单细胞多组学和成像、人工智能和机器学习、基于患者的实验性疾病模型，为未来发展提供指导。在研发赋能方面，ELIXIR 为研发数字基础设施提供了一个协调、灵活的生态环境，跨越传统赋能边界由学术界向产业界开放赋能。ELIXIR 通过向欧盟地区的生物医药中小创新企业开放平台能力，拉动欧洲生命健康产业和数字产业协同发展。ELIXIR 允许合作伙伴如其他研究基础设施、国家资源、机构档案等利用现有平台资源，连接和互操作 ELIXIR 平台资源，并在 ELIXIR 品牌保证下提供服务。在维稳资金方面，ELIXIR 采用成员国会员制度，除却“欧洲地平线”计划的资金供给，按参与国家的国民收入净额 (NNI) 收取相应比例的会费，提供技术预算用于支持和连接国家资助的生物信息资源的共享服务，形成与研发活动相适应的资金流正反馈。在数据共享方面，欧盟《通用数据保护条例》(GDPR) 为个人健康数据安全使用提供保障^[25]，EMBL 对照 GDPR 适应性调整制定了第 68 号内部文件 (Internal Policy No. 68 on General Data Protection) 作为 ELIXIR 平台的数据保护策略依据。在知识产权方面，“欧洲地平线”计

划在确保知识产权的前提下要求免费开放科研成果出版物，遵循“尽可能开放，必要时关闭”的治理原则实施数据管理计划，通过实现“可查找、可访问、可互操作和可重用”的方式开放获取研究数据，建立针对开放科学研究人员的奖励制度。

四、我国生物医药研发数字基础设施建设与发展现状

我国生物医药研发数字基础设施的建立和发展在早期以主题式集成建设模式为主。以数据资源为核心的生物医药研发数字基础设施体系已基本建立，并且正在从数据驱动向计算驱动新模式转型升级。

(一) 基本现状

早在“十二五”“十三五”时期，我国集中建立了一批主题式集成基础设施，中国科学院 40 个研究所的 72 家生物资源库馆是重要的资源型平台，包括国家化合物样品库、国家细胞资源库、生物种质类器官库和国家实验动物中心等生物种质与实验材料资源库等。以中国科学院为代表的研究机构承担建立了多个集成的生物医药数据资源型平台和主题研究平台^[26]，在科学数据层面包括国家人口健康科学数据中心 (NPHDC)、国家基因组科学数据中心 (NGDC)、中国国家数字图书馆 (NDLC)、上海国家生物医学大数据平台 (BMDC) 等。以综合性国家科学中心建设为契机，综合集成的生物医药研发数字基础设施发展态势开始显现。国家科学数据中心汇集了中国科学院科学数据中心体系下国家人口与健康科学数据共享平台、国家基因组科学数据中心等与生命科学和健康医疗领域相关的研究平台。典型的如生物医药研发计算平台的集成和集群建设处于起步阶段，先行创新试点有中国国家网格 (CNGrid)、CSTCloud 等。以中国科学院科学中心体系为代表，我国关于生物医药研发数字基础设施的建设、集成和集群的步伐与组织机构的整体组织架构保持高度一致。不足的是，由于知识产权和商业秘密等原因，国内与疾病研究、药物研究相关的主题式开放集成平台基础设施和大型研究队列较少。

“十四五”时期，生物医药产业数字化发展相关文件密集发布。2021 年 12 月，《“十四五”生物经济发展规划》推动生物信息产业发展，建立生物

技术与信息技术融合应用工程, 聚焦信息技术支撑新药研制、人工智能技术辅助诊疗和远程医疗服务。同时, 《“十四五”生物医药工业发展规划》发布, 提出医药产业化技术攻关工程, 强调生物医药技术和医疗器械技术中信息化、智能化的渗透与应用。2022年11月, 《“十四五”全民健康信息化规划》提出, 推进数字健康融合创新发展体系, 构建数字健康科技创新体系, 集约建设信息化基础设施支撑体系。面向基础设施提出“全民健康信息新基建强化工程”, 全面推进医疗卫生机构信息化建设提档升级, 鼓励各地因地制宜构建全民健康基础设施云, 推动数字健康新型基础设施建设, 全方位提升卫生健康信息化基础设施水平; 在医疗健康数据层面包括“1+5+X”健康医疗大数据平台体系、国家-省级区域卫生信息平台体系。总体来看, 我国生物医药研发体系在数字基础设施的互联互通和融合建设方面仍处于起步阶段(见表2)。

(二) 发展差距

一是国际影响力和权威性方面。我国生物医药研发数字基础设施建设起步相对较晚, 现有研发数字基础设施在国际上的权威性和影响力及行业认可度与美国等领先国家存在较大差距。部分数据库和知识库系统乃至生物医药研发数字基础设施平台的建设、运营和管理依赖于国外经验、产品供应和知识产权输送。

二是发展模式设计和效能方面。在欧盟、美国的战略中, 研发基础设施或直接作为研发项目, 或作为研究项目的子建设项和成果验收项, 通过发起

大规模研究队列计划配套建设资源型研发基础设施进一步壮大基础设施能力域、规模和影响力, 并同步加速基础设施整合建设。我国的科技创新转化长期以文章发表为重要考核指标, 必要的平台建设、连接建立或平台整合建设力度不强。对技术创新的基础设施力量 and 市场需求在生物医药整体研发链条中容易忽视, 在源头基础设施建设和使能上应发挥更大的联系作用和合力作用。生物医学大数据平台、健康医疗大数据平台及生物样本库间之间缺失广泛协同连接的机制和规范。

三是可持续建设方面。欧盟ELIXIR基础设施建设充分满足了学术研究和产业经济发展需要, 通过研究人员免费使用、以保障知识产权为前提, 鼓励开放共享、基于透明和监管支持企业共开发等系列措施, 打造了较为稳定和可持续的发展模式。我国建设生物医药研发数字基础设施面临数据碎片化、数据库建设低水平、建设标准多样化等问题, 在资源储备方面尤其是生物医药研发所需工具资源的自主研发产品较少、底层技术架构国产率低。生物医药产业以数据要素为核心的资产化生产分配、利益分配和再分配机制未建立, 生物医药数据价值创造的链条不畅致使整个领域的工程化受阻。

四是管理组织架构建设方面。欧盟基于灵活的实验室体系通过权威引领和单一管理部门引领平台集群建设不断在积聚能量方面形成更大势能差, 美国在NIH统筹下基础设施使能高度集中。相较美国的集中制管理和欧盟的协调式管理, 我国生物医药研发数字基础设施建设、运营监管仍处于“多头管理、分管分治”状态, 分散了平台能量。

表2 我国生物医药研发基础设施平台体系梳理

领域体系	建设单位	研发数字基础设施	综合集成门户
生物医药科学知识体系(中国科学技术研究院体系)	中国科学院基因组科学数据中心、中国科学院脑科学数据中心、中国科学院微生物科学数据中心、中国科学院干细胞与再生医学科学数据中心、中国科学院广州生物医药与健康研究院科学数据中心、中国科学院上海有机化学研究所	国家生物信息中心(www.cncb.ac.cn) 国家基因组科学数据中心(ngdc.cncb.ac.cn) 国家人口健康科学数据中心(www.ncmi.cn) 国家化合物样品库(www.cncl.org.cn) 国家微生物科学数据中心(nmdc.cn) 干细胞与再生医学数据中心(dscrcn) CNGrid(www.cngrid.org/) CSTCloud(www.cstcloud.cn)等	科学数据中心 (www.casdc.cn)
医疗健康领域平台体系(“1+5+X”健康医疗大数据平台体系)	国家健康医疗大数据研究院、国家区域医疗中心、省级健康医疗大数据中心	国家健康医疗大数据中心(规划中)	—

（三）主要挑战

一是顶层设计挑战。随着研究集成平台数量和种类的不断丰富，具有创新能力的整合集群生态成为当下生物医药研发基础设施建设的重要方向。新的交叉研究领域的出现将带来新的数据资源空间，量子计算等新技术的发展将带来新的计算能力，该趋势下科学融合迭代研发基础设施建设可认为是一个永恒的议题。理想的模式是通过科学建立顶层框架，从上至下有序引导生物医药研发基础设施在数字空间内不断集成、整合、融合发展，形成丰富规整的资源管线，规模化装载至具有国际影响力的创新数字基础设施平台并持续维护。通过提供一个集中的数字平台或是集中的web入口连接所有资源作为基础设施，支持研发人员工程化开展并完成关乎生命科学、人类健康和产业化的研究任务^[27]。新兴能力基础设施由下而上进行孵化和规模化发展，并与整体框架相协同。针对面向新时期和未来的生物医药研发数字基础设施进行顶层设计是一项充满挑战的任务。

二是研发数字基础设施管理和治理挑战。要建立一个完整且必要的支持科学研究能力的、连贯的基础设施存在诸多挑战，包括平台互操作能力、平台合力建设和可持续运营建设。在整合模式下，解决大量数据源或工具源的独立和异构问题，包括格式、语法和模式等^[28]的高度差异均是掣肘资源互通共享和平台集成的关键。美国、欧盟等国家和地区经过国际标准组织和研究院校等的长期研究，在发现、查询以及集成异构语法、结构、格式和生物学实体符号等方面，已研制了本体、元数据、数据集等针对生物医学和健康信息数据资源的系列标准，通过共享标识符、丰富的元数据、基于本体的工具、基于可扩展标记语言（XML）格式和应用程序编程接口（API）的灵活交换系统来解决互操作性问题^[29]。近年来，基于标准化的信息和平台互操作能力提升工作在国内开始获得高度关注，但相关标准制定和标准化方面的工作推进仍然相对落后，针对生物医药研发信息系统的底层标准化能力有待提升。

五、对策建议

生物医药研发数字基础设施是数字时代生物医

药产业面向创新驱动发展转型的创新基础设施，担负着支撑创新型国家“四梁八柱”的重任。全球以美国、中国、欧盟等为代表已在部分前沿研究领域建立了基于生物医药研究型集成平台群。未来，有战略布局基础和平台资源基础的国家将更有先发优势承建创新研发基础设施平台，平台整合带来的规模效应或将引发甚至进一步加剧全球生物医药领域创新的国家垄断行为。对我国而言，建立中国特色的生物医药研发基础设施集群化生态迫在眉睫。

（一）加强生物医药研发数字基础设施顶层设计

生物医药研发数字基础设施集群建设、规模化建设应采取生物医药研究型集成平台“整合式”发展和“主题式”并重发展策略，注重前沿关键领域研究的数字化建设和数据资产积累，注重高质量数据资源、研发工具资源集成建设和自主创新建设，最大化发挥数据战略资源价值。结合《开放科学建议书》内容，加快以“高质量网络”为关键支撑，以“数据资源、算法框架、大模型、算力资源”为核心能力要素，以“开放平台”为主要赋能载体，建设能提供公共智能化服务的研发数字基础设施具有重要意义。

（二）加强生物医学和健康医疗资源的统筹编排

强化生命科学和健康医疗领域线性贯通的知识组织和统一资源编排。统筹国内领域平台，建立生物医药全程全域数据资源目录及关联图谱，对生物医药数据库内资源进行深度关联。建立生物医药平台核心资源数字化标准体系，推动数据和知识资源的标准化治理，以数据标准为抓手推动生物医药数据资源协同管理工作，提升数据资源的继承性、一致性和连贯性，通过信息标准推动平台互联互通和协作能力。应重视数字化标准及其标准化工具的开发应用，面对当前国内数据平台林立、各类资源错落分布的状况，标准化的作用是举足轻重的，亟需上升至战略高度。

（三）加强单一个体平台集群化整合建设和治理

依照当前的快速发展态势，我们或许永远无法精确衡量或满足研究设施的不断扩大需求。因此，更好的选择是开放合作，加强生物医药核心领域研究公共数字基础设施建设。同时，采取“1+n+X”

模式逐步开展平台连接工作,推动各个生物医药自主创新平台向集群化的“共同体”方向发展。整合组建与重大疾病相关的临床表型数据、影像数据、治疗预后数据等疾病数据库,组建人体基因组、转录组、蛋白质组、代谢组、宏基因组等多组学数据库,药物分子结构、药物毒性测试、临床试验数据、真实世界数据等药物研发数据库等。建成后纳入平台集群统一管理,完善全生命周期管理,全面提升开放共享水平和运行效率。推动平台云联合以连接有助于资源池的不同设施,并节省投资成本。

(四) 加强开放平台能力建设和运营保障

加强国内平台能力建设,引入生物医药领域大模型作为新的能力资源池,壮大生物医药研发能力资源池供给。加快生物医药多模态智能大模型开发、可信的生物医药数据和知识的语义理解及推理技术研究,探索大模型在重大疾病、传染病和罕见病预防诊疗中的研究与应用。在保障平台高质量和可持续运营方面,要尽快建立并完善生物医药研发数据共享红线清单或数据目录,建立数据要素质量和安全“预认证”机制。完善研发数字资源知识产权法律法规;建立分层次的数据生产分配和价值分配体系,多措并举消弭共享意愿问题,激发价值数据所有人共享意愿。加大配套政策激励力度,充分发挥市场作用,鼓励资源所有人参与平台能力建设和合作运营。

利益冲突声明

本文作者在此声明彼此之间不存在任何利益冲突或财务冲突。

Received date: September 13, 2023; **Revised date:** October 15, 2023

Corresponding author: Li Lanjuan is a professor from State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, the First Affiliated Hospital, Zhejiang University School of Medicine, and a member of Chinese Academy of Engineering. Her major research fields include Artificial liver, infectious diseases and digital medicine. E-mail: ljli@zju.edu.cn

Li Xiaokun is a professor from the Wenzhou Medical University, and a member of Chinese Academy of Engineering. His major research field is basic theoretical research and new drug research of protein drugs represented by growth factors. E-mail: xiaokunli@163.net

Tan Weihong is a professor from Hangzhou Institute of Medicine, Chinese Academy of Sciences, and a member of Chinese Academy of Science. His major research fields include Biochemistry, biochemical analysis, functional nucleic acids, and molecular medicine. E-mail: tan@hnu.edu.cn

Funding project: Chinese Academy of Engineering project “Strategic Research on the Development of Integration Platform within Artificial Intelligence and Big Data for the Biomedical Industry” (2022-DFZD-28)

参考文献

- [1] Mirowski P. The future(s) of open science [J]. *Social Studies of Science*, 2018, 48(2): 171–203.
- [2] Fecher B, Kahn R, Sokolovska N, et al. Making a research infrastructure: Conditions and strategies to transform a service into an infrastructure [J]. *Science and Public Policy*, 2021, 48(4): 499–507.
- [3] Zhang L L, Li J H, Paul F, et al. Research e-infrastructures for open science: The national example of CSTCloud in China [J]. *Data Intelligence*, 2023, 5 (2): 355–369.
- [4] Zhao Z, Huang J. Analysis of information resource construction mode of open science infrastructures [J]. *Library Development*, 2021, 44(3): 46–55.
- [5] UNESCO. Recommendation on open science [EB/OL]. (2021-11-26) [2023-08-22]. <http://www.unesco.org/en/articles/unesco-sets-ambitious-international-standards-open-science>.
- [6] Hardisty A, Roberts D, The Biodiversity Informatics Community. A decadal view of biodiversity informatics: Challenges and priorities [J]. *BMC Ecology*, 2013, 13: 16.
- [7] National Academies of Sciences, Engineering and Medicine, Policy and Global Affairs, et al. Open science by design: Realizing a vision for 21st century research [M]. Washington DC: The National Academies Press, 2018.
- [8] 余辉, 梁镇涛, 鄢宇晨. 多来源多模态数据融合与集成研究进展 [J]. *情报理论与实践*, 2020, 43(11): 169–178.
- [9] Yu H, Liang Z T, Yan Y C. Review on multi-source and multi-modal data fusion and integration [J]. *Information Studies: Theory & Application*, 2020, 43(11): 169–178.
- [10] 周小林, 李力, 杨云, 等. 欧盟大型研究基础设施路线图的经验及对中国大科学监测评估的启示 [J]. *中国科技论坛*, 2020 (1): 181–188.
- [11] Zhou X L, Li L, Yang Y, et al. Experience in the EU large research infrastructure roadmap and implications for big science monitoring and evaluation in China [J]. *The China Science and Technology Forum*, 2020 (1): 181–188.
- [12] Ribes D. The kernel of a research infrastructure [C]. New York: Association for Computing Machinery, 2014.
- [13] Ramaprasad A, Valenta A L, Brooks I S, et al. Biomedical informatics infrastructure [J]. *Urban Economics & Regional Studies eJournal*, 2007: 1305712.
- [14] Davidson S M, Heineke J. Toward an effective strategy for the diffusion and use of clinical information systems [J]. *Journal of the American Medical Informatics Association*, 2007, 14 (3): 361–367.
- [15] Lauer D K B, Smith A, Blomberg D N, et al. Open data: A driving force for innovation in the life sciences [EB/OL]. (2021-08-19) [2023-08-22]. <https://fl1000research.com/documents/10-828>.
- [16] U.S. National Library of Medicine. A platform for biomedical discovery and data-powered health strategic plan 2017–2027 [EB/OL]. (2018-03-06) [2023-08-22]. https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport2017_2027.pdf.

- [15] National Institutes of Health. NIH-Wide strategic plan for fiscal years 2021—2025 [EB/OL]. (2021-07-30)[2023-08-22]. <https://www.nih.gov/sites/default/files/about-nih/strategic-plan-fy2021-2025-508.pdf>.
- [16] National Institutes of Health. About the STRIDES initiative [EB/OL]. (2021-07-30)[2022-11-01]. <https://datascience.nih.gov/strides>.
- [17] 苏燕, 李伟, 李祯祺, 等. 美国生物大数据战略举措及其对我国的启示 [J]. 中华医学图书情报杂志, 2020, 29(10): 32–37.
Su Y, Li W, Li Z Q, et al. Strategic measures taken by USA for biological big data and their enlightenments [J]. Chinese Journal of Medical Library and Information Science, 2020, 29(10): 32–37.
- [18] National Institutes of Health. The NIH BRAIN initiative: An overview [EB/OL]. (2020-10-05) [2022-11-01]. <https://www.nlm.nih.gov/news/events/2020/townhall/the-nih-brain-initiative-an-overview>.
- [19] Collins F S, Schwetz T A, Tabak L A, et al. ARPA-H: Accelerating biomedical breakthroughs [J]. Science, 2021, 373(6551): 165–167.
- [20] Hillson N, Caddick M, Cai Y, et al. Building a global alliance of biofoundries [J]. Nature Communication, 2019: 2040.
- [21] Shaping Europe’s digital future: Commission presents strategies for data and Artificial Intelligence [EB/OL]. (2020-02-19)[2022-11-01]. https://futurium.ec.europa.eu/sites/default/files/2020-08/Shaping_Europe_s_digital_future_Commission_presents_strategies_for_data_and_Artificial_Intelligence.pdf.
- [22] Harrow J, Drysdale R, Smith A, et al. ELIXIR: Providing a sustainable infrastructure for life science data at European scale [J]. Bioinformatics, 2021, 37(16): 2506–2511.
- [23] Uuropean commission CORDISEU research results [EB/OL]. (2021-03-15)[2023-10-12]. <https://cordis.europa.eu/projects/en>.
- [24] Horizon Europe’s first strategic plan 2021—2024: Commission sets research and innovation priorities for a sustainable future [EB/OL]. (2021-03-15)[2022-11-01]. https://commission.europa.eu/system/files/2021-09/ec_rtd_horizon-europe-strategic-plan-2021-24.pdf.
- [25] 俞胜杰. 《通用数据保护条例》中的域外管辖问题研究 [D]. 上海: 华东政法大学(博士学位论文), 2020.
Yu S J. Research on extraterritorial jurisdiction in the *General data protection regulations* [D]. Shanghai: East China University of Political Science and Law (Doctoral dissertation), 2020.
- [26] 樊代明. 生物医学大数据是重要战略资源 [J]. 科学新闻, 2019 (6): 34.
Fan D M. Biomedical big data is an important strategic resource [J]. Science News, 2019 (6): 34.
- [27] Zhang L L, Li J H, Uhlir P F, et al. Research e-infrastructures for open science: The national example of CSTCloud in China [J]. Data Intelligence, 2023, 5(2): 355–369.
- [28] Kamdar M R, Fernáandez J D, Polleres A, et al. Enabling web-scale data integration in biomedicine through linked open data [J]. npj Digital Medicine, 2019, 2(1): 1–14.
- [29] Laufs D, Peters M, Schultz C. Data platforms for open life sciences—A systematic analysis of management instruments [J]. PLoS One, 2022, 17(10): e0276204.