欧美生物医学科学数据中心建设及启示*

吴思竹[1] 王安然[1] 修晓蕾[1] 钱庆[1] 周伟[2]
(1. 中国医学科学院医学信息研究所，北京 100020；2. 国家人口健康科学数据中心，北京 100005)

摘要：本文全面分析了具有重要国际影响力的美国国家生物技术信息中心和欧洲生物信息学研究所在生物医学数据资源建设、工具利用和共享服务等方面的发展经验，并从经费人员基础保障、基础设施、资源建设、技术研发、标准规范和用户服务六方面为我国开展生物医学科学数据中心建设提出启示和建议。

关键词：生物医学；科学数据中心；数据管理；数据共享
中图分类号：G203 DOI：10.3772/j.issn.1673-2286.2022.04.001
引文格式：吴思竹，王安然，修晓蕾，等. 欧美生物医学科学数据中心建设及启示[J]. 数字图书馆论坛，2022(4)：2-10.

1 欧美典型领域科学数据中心概述

1.1 美国国家生物技术信息中心

美国国家生物技术信息中心（National Center for Biotechnology Information，NCBI）是美国国立卫生研究院国家医学图书馆（National Library of Medicine，NLM）的下属中心，成立于1988年。NCBI的定位和重要使命是研发前沿信息技术，帮助理解和控制健康以及疾病的基本分子和遗传过程，创建存储和分析有关分子生物学、生物化学和遗传学相关知识的自动化系统，促进研究和医学界开展数据库和软件应用，协调国内和国际范围开展收集生物技术信息、研究生物重要分子结构和功能分析的先进的计算机处理方法[1]。

NCBI的组织架构由科学顾问委员会、计算生物学部、信息工程部和信息资源部组成。计算生物学部是NCBI的主要研究部门，包括9个研究小组，分别负责染色质结构生物信息学、序列比对、蛋白质和基因组进化、进化基因组学、生物医学文本挖掘、表观遗传和基于序列的基因调控机制、生物系统建模和计算分析、统计计算生物学和定量分子生物物理学研究。信息工程部主要负责设计和构建NCBI的软件系统和数据库。信息资源部负责规划、指导和管理NCBI的技术和网络运营，为内外部用户提供技术支持、咨询和指导。2021年NCBI的全职员工有273人，占NLM全职员工数量（647人）的42%，主要由研究人员、跟踪调查人员、软件工程师、博士后和学生等多层次人员组成。NCBI的研究和服务主要由NLM内部计划（Intramural Programs）的经费支持，该经费占拨付给NLM的美国总统预算经费的80%，主要支持NLM开展前沿计算健康信息学研究，开发先进的生物医学信息系统、标准和研究工具，开展生物医学数据采集、存储、传播，以及提供高质量的信息服务。2017—2021年，NLM的这项计划经费总额一直保持在3亿~4亿美元。通过公开的NLM预算收支可获知2008年以前的NCBI获得的支持经费为7 000万美元左右（占NLM全部预算经费的23%~25%）。虽然之后财年预算没有公开NCBI的支持经费，但从2022年NCBI主任和NLM科学数据资源副主任职位招聘的

* 本研究得到中国医学医科院医学与健康科技创新工程项目“人口健康科学大数据智能管理与高效利用技术体系建设”（编号：2021-I2M-1-057）资助。

启事中可获知，该职位可以有权支配1.5亿美元的经费预算和领导NLM的近700名员工[2]。

1.2 欧洲生物信息学研究所

欧洲生物信息学研究所（European Bioinformatics Institute，EBI）是欧洲分子生物学实验室（European Molecular Biology Laboratory，EMBL）的一部分，它的定位和使命包括五方面：为科学界建立和维护生物数据库，提供科学服务和培训，开展生物信息学基础研究，向工业界传播前沿技术及协调供应欧洲生物数据[3]。

EBI的组织架构包括理事会、战略管理委员会、科学顾问委员会、研究组、服务组、技术组、培训组等多个部门及团队。研究组包括4个小组，分别负责系统和数学生物学、蛋白质结构和化学、基因组和功能基因组学研究。服务组主要负责基因、基因组和变异，分子图谱，蛋白质组和蛋白质家族，分子系统，分子和细胞结构，化学物质等数据和文献服务。技术组主要负责系统应用、系统基础架构建设、网络和软件开发与运维。EBI的人员队伍具有多元化和多学科特点，由来自78个国家的850多人构成，包括研究人员、编外人员、学生和访问人员。2020年，EBI的全职工作人员有697名，其中在职员工618人，博士后42人，博士生37人。EBI的大部分经费来自EMBL的20多个成员国的政府公共经费，还有部分来自国际合作资助经费，国际合作的主要资助者包括英国研究与创新署（UK Research and Innovation）、英国的生物技术和生物科学研究委员会（Biotechnology and Biological Sciences Research Council）、欧盟委员会（European Commission）、美国国立卫生研究院（National Institutes of Health，NIH）和惠康信托基金会（Wellcome Trust）等。在支撑经费上，EBI 2017—2020年的年支出经费在8 000万欧元左右，2020年支出8270万欧元，其中数据资源建设和服务占55%，基础维护、技术开发和IT基础设施支撑占15%，研究占14%，培训占6%，管理和房地产成本占10%。

NCBI和EBI作为世界上较大的生物医学科学数据中心，其工作分别在NLM和EMBL领导下开展，并高度对齐NLM和EMBL的最新战略计划。NLM在2018年发布《生物医学发现和数据驱动健康平台：2017—2027年国家医学图书馆战略计划》（*A Platform for Biomedical Discovery and Data-Powered Health National Library of Medicine Strategic Plan 2017-2027*），EMBL在2018年发布《EMBL计划2017—2021：数字生物学》（*EMBL Programme 2017-2021：Digital Biology*），在2022年发布《EMBL计划2022—2026：从分子到生态系统》（*EMBL Programme 2022-2026：Molecules to Ecosystems*），二者战略的核心主题均将开放科学、基础设施能力的提高、数据科学研究与推动、服务培训增强、数据应用创新与转化等作为重要目标。NCBI和EBI面向战略目标承担重要使命，包括数据资源建设维护，关键技术工具研发，面向世界范围的专业知识服务、教育和培训等。

2 建设模式和主要进展

2.1 数据资源体系建设

2.1.1 重视开展高质量专业数据库建设

NCBI和EBI均包括多来源数据，主要分为三类：①科研人员提交的数据资源，如SRA、PRIDE等收集的研究人员提交的数据；②与数据供应商和研究联盟国家合作的数据资源，例如Genbank和ENA均是国际核苷酸序列数据库协作体（International Nucleotide Sequence Database Collaboration，INSDC）的一部分，通过遵守公共数据交换标准与DDBJ进行结构统一的核苷酸序列数据交换共享；EBI与人工智能公司DeepMind合作建设的AlphaFold蛋白质结构数据库；③由数据中心专业人员加工审编的数据资源，如PubChem BioAssay、UniGene等。NCBI和EBI的数据资源建设模式以专业数据库建设为主，重视资源的广度和深度建设。广度建设包括：高通量测序原始数据的长期收集，领域文献资源建设，涵盖基因表达、基因组、蛋白质、结构、系统、化学物质、临床等多类主题专业数据库建设，关联不同类型数据的综合性数据库建设，以及本体词表等知识组织系统建设。数据资源深度建设主要结合不同主题特点收集数据或在已有原始数据基础上，通过生物医学专业人员或权威领域专家的注释、审编、集成和二次分析加工等形成具有专门用途的高质量特色主题数据库。同时，数据库建设结合Web浏览器、图形可视化分析、人机交互等技术，提升数据的展示效果、可理解性和易用性。这些数据库资源由大量高水平计算机和生物医学专家团队进行长期建设、更

新维护和优化升级，数据具有较高质量和时效性。

2.1.2 数据库资源规模大且数量增长迅速

NCBI和EBI汇集的各类型数据增长迅速，每年会在*Nucleic Acids Research*上发布其资源建设进展。根据NCBI在*Nucleic Acids Research*上发布的2021年的35个数据库的列表，本文汇总整理了这些数据库在2018—2021年的资源数量年均增长率情况[4-7]。如图1所示，68%的数据库的数据量呈逐年递增趋势，31%的数据库的年均增长率超过15%，其中Assembly的年增均长率达到77.28%，SRA、Identical Protein Groups等的年均增长率均在30%以上。2021年底SRA的数据规模已超过36PB。

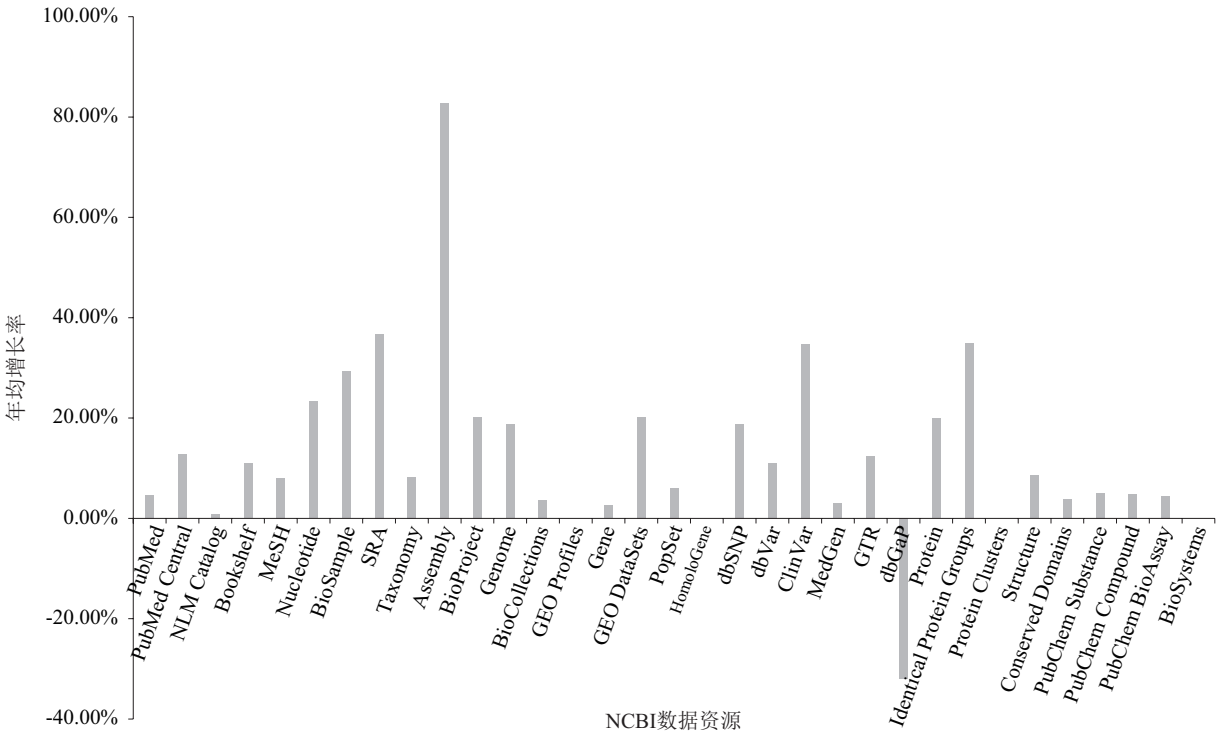

图1 NCBI数据资源年均增长率（2018—2021年）

EBI的2020年度报告显示其人类基因组和表型组数据存储量增长超过50%，其中，包括电子显微镜数据在内的成像数据存储量超过之前所有年度存储量的120%，而电子冷冻显微镜数据增长164%[8]。截至2021年EBI网站发布数据库资源数量为43个，其存储数据规模也已达390PB。

2.1.3 制定多类数据标准规范数据共享

NCBI和EBI在资源建设方面均非常重视数据标准建设和应用。本文通过FairSharing网站检索创建机构包含NCBI和EBI，并且状态为“Ready”（标识已发布应用）的数据标准，汇总分析两个数据中心建立的数据标准现状[9]。经过去重，共检索得到45个标准。其中，NCBI有10项标准（如A Gold Path format、GenBank Sequence Format、Cell Behavior Ontology等），EBI有37项标准，有2个格式交换标准是由NCBI和EBI共同参与创建的（INSD Sequence Record XML和DDBJ/ENA/GenBank Feature Table，用于INSDC联盟进行核苷酸数据交换共享）。其他标准类型有20项数据格式标准（如EBI BioSamples JSON Format、ENA Sequence Flat File Format、OmicsDI XML Format等）、14项术语标准（如Systems Biology Ontology、Human Phenotype Ontology等）、8项数据集元数据标准（如FAANG Metadata Experiment Specification Standard和Minimum Information about Plant Phenotyping Experiment等）和3项数据分类标准（如UniProt Taxonomy等）。通过综合分析，本文发现两个数据中

心的标准建设工作聚焦在数据的规范化表示和描述方面，重点解决多种类型的生物医学数据在提交、存储、下载、计算和交换中的文件格式问题。其次是语义表达差异性问题和数据描述的结构化与规范性问题。NCBI和EBI也积极开展标准的应用和服务。它们在用户数据提交时要求使用标准数据格式，如核酸序列存储使用FASTA格式、存储核酸序列和测序质量信息用FASTQ格式、存储序列比对结果用SAM/BAM格式、存储变异文件用VCF格式、保存遗传特征描述用GFF3格式等。NCBI提供MeSH主题词表和NCBI Taxonomy的浏览查询服务。EBI不仅提供多种本体资源并开展统一本体查询服务，创建的Ontology Lookup Service网站收录了近280个本体、700万条术语和50万个实例[10]。同时，EBI也参与了全球基因组学与健康联盟（Global Alliance for Genomics and Health），积极推动基因组研究和医疗健康数据共享的国际政策和标准，提高其对临床研究社区不断变化的需求的服务水平。NCBI和EBI创建和参与制定的多类标准也在领域中被广泛实施和复用，开展了良好的应用实践。

2.2 关键技术工具研发

2.2.1 扩展云基础设施提升存储和服务

生物医学数据的指数级增长和数据密集型科学研究需求的日益迫切，NCBI和EBI积极开展EB级海量数据处理、存储、计算和服务的解决方案探索。2019年，EBI获得英国研究与创新署的4 500万英镑的投资，用于提高其基础设施建设能力。2020年开始，NCBI和EBI在加强基础设施建设的同时积极探索应用云环境加速促进其研究创新和满足全球化用户服务需求。早在2013年，欧洲就组建了ELIXIR（http://www.elixir-europe.org/）负责协调欧洲的数据、工具、云存储、超算及培训等资源以建立一个可持续的泛欧生物信息研究基础设施。EBI是ELIXIR项目重要节点，已建有庞大的技术基础设施，包括虚拟化环境、高性能计算集群和近440PB的存储资源。为了满足快速增长的运营需求，EBI先后与Google Cloud和Amazon Web Services（AWS）建立战略合作，实施混合云及多云战略。公有云主要使用Google Cloud和Amazon Web Services，私有云使用Embassy Cloud和欧洲开放科学云（EOSC）。依托私有云开展大规模的国际合作，如泛癌症全基因组分析和Tara Oceans等；依托公有云，EBI已开展Human CellAtlas、Idenitifers.org、Ensembl镜像等项目，为研究人员提供各类分析计算工具和服务。

NCBI基于NLM建立的发现、实验和可持续性科学技术研究基础设施（STEARDS）计划，先后开展了与Google Cloud、Amazon Web Services和Windows Azure的合作，利用云平台进行SRA、COVID-19基因组序列数据集、BLAST数据库和PMC等多个数据库的数据托管和维护，并利用云环境支撑大规模数据的传输、分析、计算和科研协作。

2.2.2 利用高速工具提高数据获取效率

由于生物医学数据规模大且用户利用率高，NCBI和EBI研发以及应用能够支持高性能的大规模数据上传、下载和集成检索工具以提高数据获取效率。两个数据中心均提供IBM Aspera软件作为大数据传输的解决方案。Aspera基于IBM FASP传输协议，传输速度远高于FTP，内置AES-128加密算法，支持传输加密、落地解密和断点续传，并提供浏览器插件、客户端和命令行等多种使用方式。除Aspera以外，针对不同类型的资源，NCBI还提供NCBI E-Utilities、SRA Toolkit、GEO2R等工具，以及提供API接口等服务方式支持大体量特定类型数据资源的下载。

在数据集成检索方面，NCBI和EBI分别创建了高性能数据库集成检索系统NCBI Entrez和EBI Search，并提供统一检索结果展示和分类筛选页面。NCBI Entrez支持NCBI六大类38种数据库资源检索。用户可以通过浏览器访问NCBI Entrez，也可以使用E-Utilities通过程序接口和参数设置进行按需数据调用[11]。EBI Search与EBI的数据资源同步，57.14版本已索引了近48亿个条目。它底层基于Apache Lucene建设，利用Carrot2实现查询结果的优化[12-13]。EBI Search提供Web浏览器和RESTful接口访问[14]。NCBI Entrez和EBI Search最大的共性特点在于均能够支持数据库记录内和数据库记录之间的广泛链接和交叉引用，可以帮助用户发现数据关联和扩展应用。如NCBI Entrez包含两种数据关联：一种是通过相似度等计算获得的，如基于BLAST相似性搜索发现相关序列；另一种是记录数据本身存在的关系或跨库资源类型间的关联，如PubMed的论文摘要与PMC全文的关联，蛋白质序列与其编码DNA序列或发表它的论文之间的关联等[15]。

2.2.3 研发多类数据处理分析工具软件

结合不同类型生物医学数据的数量、特点和用途等，NCBI和EBI研究开发了大量文献查询注释（PubMed Clinical Queries、Annotation Platform和Europe PMC Grant Finder等），蛋白质、核酸和DNA序列比对，包括序列相似性搜索（BLAST、Simple Phylogeny和FASTA等）、多序列比对（Kalign、T-COFFEE、CLUSTALW2和COBALT等）、双序列比对（Clustal Omega和PSI-BLAST等），分子进化和系统发育树构建（CDTree和Lifemap等），结构分析与可视化显示（Cn3D、MapViewer、Nightingale和Protvista等）相关工具。门户展示的工具数近50个，满足不同用户的数据处理、分析和挖掘需求。NCBI的研究小组也在不断优化数据库搜索、序列比对、基因组分析、图像注释、蛋白质结构和功能预测等方面的算法和工具。为提高共享应用，数据中心网站均提供多种工具服务形式，包括云服务、网络版、工具包下载、开发接口等，允许用户通过Web、RESTful API和命令行等创建自定义数据集，并支持结构化文件下载。同时，其也为具有开发能力的用户提供开发接口、工具包或开源代码。此外，两个数据中心均在GitHub.com发布了部分数据资源和工具代码，提供用户开放下载和获取。

2.3 多元数据共享服务

2.3.1 保障数据开放共享和安全利用

NCBI和EBI的生物医学数据资源和工具在开放获取政策支持下，开放和开源程度极高，研究人员可访问大部分数据，并从网站公开获取和下载。开放的数据资源及工具多基于CC-By、CC0、Apache 2.0等开放协议。研究人员提交到NCBI的人类基因组数据必须遵循《美国国立卫生研究院的基因数据共享政策》（*Genomic Data Sharing Policy*），依照《美国联邦受试者保护通则》（*Federal Policy for the Protection of Human Subjects*）和《美国健康保险可携性和责任法案》（*Health Insurance Portability and Accountability Act*）隐私条例标准进行识别化处理。对于涉及受试者个人信息和人类遗传资源的数据，NCBI严格实施受控访问。例如，在数据提交到GEO、SRA等数据库之前，研究人员需要在原始数据清洗和质控前完成dbGap注册。dbGap是提供遗传关联研究、甲基化研究和其他高风险个体水平数据的数据库，其仅供研究目的使用，数据的访问需要通过数据访问委员会（Data Access Committee，DAC）审核和授权。

而EBI的开放数据遵循EMBL的开放获取政策，促进数据的广泛和免费获取。同样，其也对潜在可识别的人类基因数据访问进行严格控制，如存储研究项目产生的所有类型的个人可识别遗传和表型数据的欧洲基因组表型数据库（EGA）有7 713个数据集，数据使用仅面向特定研究用途或研究人员发布。这些数据集中包含25个受限数据集，使用需要向DAC申请并提交研究方案，审批授权后方能访问。

2.3.2 重视数据领域专业知识技能培训

两大数据中心非常重视面向各层次的科学研究者（包括领域科学家、临床医生、技术开发人员和学生等）提供数据和工具专业知识技能的指导和培训，提供较为全面的数据提交、下载、工具安装、使用流程和方法说明。如由于数据资源库种类过于繁杂，NCBI和EBI的数据汇交门户均提供数据提交引导，通过分步选择引导用户明确需求找到适合的存储数据库，降低使用门槛。此外，专业人员教育和用户培育也是两个数据中心的重要工作，NCBI创建了各种在线课程、线上和线下研讨会，提供视频、培训材料和文档，并充分利用多渠道和多媒体开展资源和工具的宣传、动态报道和与用户交互，包括YouTube、博客、社交媒体网站（FaceBook、Twitter和LinkedIn）、RSS、邮件和NCBI Insight网站等。NCBI不定期举办“编程马拉松”活动，鼓励不同背景的研究人员、开发人员和数据科学家、学生和博士后等组建团队，帮助参与者深入了解数据管理、学习编码的最佳实践和创新生物医学数据计算分析模型算法及工具。EBI也积极开展多种生物医学专业知识培训，包括现场培训、网络研讨会、虚拟课程和在线教程集合等，培训方式包括实时课程、点播课程和培训师支持。实时课程主要提供生物分析的特定领域培训，网络研讨会计划和在线教程侧重于EBI资源和工具的介绍和利用。培训通过邀请EMBL及领域或社区的专家以指导实践练习、开展小组讨论和问题交流为主。2020年，已有54.5万个独立IP用户访问了EBI的相关培训页面，通过多种数据中心提供的培训方式提高了生物医学数据素养水平[16]。

2.3.3 开展研究合作积极推动数据增值

NCBI和EBI拥有高水平的研究组，积极开展深入的生物医学信息学和数据科学研究及合作。本文通过对PumMed数据库进行检索，统计得到：NCBI在1990—2021年发表论文3 779篇，EBI在1994—2021年发表论文3 166篇。这些论文主要发表在*Nucleic Acids Research*、*Genome Biology*、*Nature*等具有影响力的期刊上，主要是报道关于序列和结构比对算法，基因组分析可视化工具，基于深度学习的生物图像解释，基础生物学突破，以及其他具有广泛性与重要性的创新算法、方法、资源和工具的相关研究和实践[17-19]。两大数据中心始终保持在国际生物医学信息学研究和应用领域的先进性和影响力。由于其开展了领域数据管理和共享的良好实践，它们也成为Springer Nature、Wiley、Elsevier等知名出版商的数据政策中所推荐的可信赖的科学数据仓储，为很多重要的学术论文发表和传播提供了可靠的数据来源、长期的数据存储和高性能计算分析等重要支持，获得用户的长期信赖。同时，NCBI和EBI也积极和多方开展研究合作，参与大型科研项目。NCBI的研究人员与NIH内的多个研究所以及众多学术界或政府开办的研究实验室保持着持续的合作。EBI的研究人员参与了人体细胞图谱（Human Cell Atlas，HCA）、OpenTargets、ICGC-ARGO、泛癌全基因组分析（Pan-Cancer Analysis of Whole Genomes）等大型研究计划。在与非学术型机构合作方面，EBI面向全球“20强”制药公司及农业食品、营养和医疗保健公司提供研究计划及专业知识交流平台，组织季度战略会议和专家研讨会等活动，并为中小企业发展和技术产品转化提供必要的数据基础设施、数据和服务，帮助其加速产品研发与创新。

3 对我国领域科学数据中心建设的启示

随着大数据、物联网和人工智能等新技术在生物医学研究中的应用，我国生物医学领域数据规模骤增，成为全球重要的生物医学数据生产国，具有丰富的民族遗传资源、家系遗传资源、典型疾病临床病例资源等重要数据资源。我国虽是数据生产大国，但数据资源利用水平低，生物医学数据资源建设和开放共享服务存在基础设施支持不足、高质量数据规模不够、数据标准化程度低、价值挖掘服务能力不足等系列问题。本文通过系统分析NCBI和EBI在资源建设、技术工具和共享服务等方面的建设和发展经验，为我国开展生物医学科学数据中心建设提出了启示和建议。

3.1 基础保障方面

（1）数据战略下的协同发展。NCBI和EBI在NLM和EMBL整体战略计划的推动下，开展数据中心建设，持续保持其在全球生物医学领域资源建设、数据管理、计算分析、数据科学教育培训等方面的领先优势。两个中心定位清晰，并与NLM及EMBL的其他部门紧密合作、优势互补、协同发展。我国生物医学科学数据中心也应积极结合国家科学数据战略，与领域/行业伙伴合作，积极补短板、强弱项，全面提升核心竞争力。

（2）提供稳定资助经费投入。NCBI和EBI均具有持续稳定的大规模经费投入，特别是政府经费。相较国外，我国数据中心建设起步晚，资助经费来源单一，不足以支撑PB级数据快速增长所带来的基础设施建设、数据存储、长期保存、平台工具研发维护、用户教育培训、人员队伍建设和管理等巨额成本，亟需国家加大对数据中心经费投入和拓展多种资助渠道。

（3）建立稳定专业人才队伍。虽然NCBI全职员工数量少于EBI，但是NCBI与NLM的研究、资源和服务结合紧密，共同开展MeSH、PubMed等多类资源和衍生工具的建设以及服务的开展。但我国数据中心的全职员工数量还远不足此。因此，我国生物医学科学数据中心亟需扩大人员队伍规模、增强团队多样性、提高团队研究水平和待遇水平。

3.2 基础设施方面

（1）加强高性能基础设施建设。NCBI和EBI均有政府支持的大规模经费用于提高中心的计算、存储等基础设施建设能力。自2019年我国科学数据汇交工作启动以来，数据PB级增长，对已有基础设施带来了严峻挑战。我国生物医学科学数据中心需要加强构建强大且可访问的数据基础设施，这对于未来几十年的生物医学科学研究发现至关重要。

（2）探索可靠的云平台解决方案。通过NCBI和EBI在私有云和公有云方面的探索，让我们看到了云平台在生物医学数据科学研究和共享服务中发挥的重要作用。其不仅可以支持用户根据自身需要访问、分析、

计算大规模生物医学数据，也可以降低数据中心对基础设施的管理和维护成本。我国生物医学科学数据中心也应积极探索私有云、公有云及混合云的数据存储和服务策略，但还需要综合考虑解决好数据的流转、存储和计算安全和监管问题。

3.3 资源建设方面

（1）扩展多种数据资源渠道。我国数据中心应深化《科学数据管理办法》贯彻落地，加强政府预算支持的科技项目的科学数据汇交管理。同时，扩展与生物医学领域研究机构及医疗行业的交流合作，通过资助合作、国际合作、协议合作等不同方式拓展数据资源创建渠道。数据中心应提高研究和服务水平，不断识别和发现新类型数据资源和开发新的服务方式。

（2）优化数据资源内容质量。结合生物医学领域数据标准和知识组织体系，加强对原始数据的归类、重组、注释、关联和整合，做好数据质量审核和控制。面向不同使用需求，研发多类主题数据库、参考数据库、整合数据库及创新型数据库和知识库。

（3）增强数据资源的FAIR化。对齐国际数据中心发展趋势，遵循FAIR原则，基于唯一标识技术、语义技术、Web浏览器技术、人工智能和可视化等技术，增强数据的可理解性、可用性、易用性和互操作性。

3.4 技术研发方面

（1）突破大规模数据处理瓶颈。我国生物医学科学数据中心迫切需要构建和使用支持大规模数据上传、下载、处理、压缩、存储、检索、质控和长期保存等系列工具，解决数据中心发展中面临的大规模数据处理和管理性能差、效率低的瓶颈问题。数据中心应通过利用人工智能、区块链、联邦学习、多方安全计算等关键技术，面向日益增长的跨组织/机构大数据协同分析和安全计算需求，研发高性能、流程化的协同分析平台以及数据挖掘模型和工具，支持数据驱动建模、模型驱动数据分析，实现生物医学数据的分析增值。通过平台和工具的建设，提高用户的数据挖掘分析效率，并保持科研过程的透明性和结果的可复制性，最终实现生物医学大数据的安全共享和跨组织协同分析的目标。

（2）研发自主可控的技术工具。NCBI和EBI研发了大量生物医学研究所必须的数据资源处理、标注、比对、分析、挖掘和预测等关键技术工具。虽然目前这些资源和工具大多数是向全球开放和开源获取的，但面对国外数据资源和技术垄断、停止更新或服务提供等情况，我国生物医学科学数据中心应当积极开展自主可控的关键技术工具研发，构建安全可靠的国产替代型工具，努力开展核心技术源头创新，提高数据中心的科技创新自强自立水平。

3.5 标准规范方面

（1）加快实用标准建设落地。数据标准对开展生物医学数据建设、管理、共享和利用起到重要指引和规范作用。NCBI和EBI积极主导了大量在生物医学领域具有重要影响的基因组和蛋白质组学等相关优质数据集的描述、数据表示和数据互操作等标准规范制定，开展了最佳实践工作。我国的生物医学科学数据中心应在结合领域发展和标准建设现状的基础上，优化完善已有生物医学数据标准规范体系，重点针对数据质控、数据分级、数据共享等方面，分阶段、有步骤地加强核心标准的研制，重点制定专业领域空白和缺失的数据标准，持续开展已有数据标准的修订和完善。

（2）参与国际数据标准制定。应积极参与国际生物医学科学数据标准规范的制定，一方面，重视国际标准的采纳、本地化和引用；另一方面，加快国家标准与国际标准的接轨，提高我国生物医学科学数据标准制定水平，提升我国国际数据标准制定的话语权。此外，应重点结合生物医学科学数据汇交、质控、整合、存储、交换和共享实践开展标准规范宣传、推广和落地监管，促进标准规范切实应用和发挥有效作用，让生物医学科学数据中心建设和发展有标可依，行之有效。

3.6 用户服务方面

（1）加强多类型用户培训。NCBI和EBI已从用户社区建设和交互反馈中获得重要改进和影响力，我国生物医学科学数据中心也应积极面向学生、研究人员、数据管理人员、企业用户等不同类型群体，开展生物医学科学数据管理和数据科学基础及专业知识培训，为用户提供生物医学科学数据管理计划、资源查找、分析挖掘、共享利用等方面的咨询和指导。

（2）有效提升用户参与度。新型冠状病毒肺炎大流行极大地促进了线上活动，数据中心应充分借助在

线视频、在线会议、微博、微信、QQ等多媒体工具，并通过举办线上、线下相结合的培训课程、数据竞赛、校外实习等使用户了解和参与生物医学科学数据管理最佳实践。

（3）促进数据和研究转化增值。数据中心应积极利用已有研究成果、数据资源以及关键技术工具，加强与领域专家、研究机构和生物医疗机构及企业的合作，参与国际项目合作。从丰富的研究和合作中，一方面积极获取学术型研究需求和经验，深化生物医领域数据管理研究；另一方面积极获取非学术型的应用需求，推动生物医学科学数据驱动的应用创新和成果转化。

4 结语

2021年联合国教科文组织正式发布《开放科学建议书》（*UNESCO Recommendation on Open Science*），标志着开放科学进入全球共识新阶段[20]。NCBI和EBI作为生物医学领域数据开放共享典型代表的数据中心已取得了较为显著的成效和影响力，本文系统梳理了其在数据资源体系建设、关键技术工具研发和多元数据共享服务等方面的进展，并基于此，探讨对我国生物医学科学数据中心发展的启示和建议，为我国生物医学领域相关数据中心在“十四五”期间一步深化数据中心资源和服务建设以及长期发展提供借鉴思路。本文研究中也还存在不足，由于两家数据中心建设成果丰富，因笔者研究精力和学科所限，在内容揭示的全面性和分析的深入性方面还存在局限，将在后续研究工作中持续完善。

参考文献

[1] NCBI. Our Mission [EB/OL]. [2021-12-03]. https://www.ncbi.nlm.nih.gov/home/about/mission.

[2] NCBI Director, NLM [EB/OL]. [2022-03-03]. https://irp.nih.gov/careers/faculty-level-scientific-careers/ncbi-director-nlm.

[3] EBI Mission [EB/OL]. [2021-12-03]. https://www.ebi.ac.uk/about.

[4] SAYERS E W, AGARWALA R, BOLTON E E, et al. Database resources of the National Center for Biotechnology Information [J]. Nucleic Acids Research, 2019, 47(D1): D23-D28.

[5] SAYERS E W, BECK J, BRISTER J R, et al. Database resources of the National Center for Biotechnology Information [J]. Nucleic Acids Research, 2020, 48(D1): D9-D16.

[6] SAYERS E W, BECK J, BOLTON E E, et al. Database resources of the National Center for Biotechnology Information [J]. Nucleic Acids Research, 2021, 49(D1): D10-D17.

[7] SAYERS E W, BOLTON E E, BRISTER J R, et al. Database resources of the National Center for Biotechnology Information [J]. Nucleic Acids Research, 2022, 50(D1): D20-D26.

[8] EMBL EBI 2020 Annual Report [EB/OL]. [2021-12-03]. https://www.embl.org/documents/wp-content/uploads/2021/07/EMBL-EBI_2020_Annual-Report.pdf.

[9] Fairsharing [EB/OL]. [2021-12-03]. https://fairsharing.org.

[10] EMBL-EBI Ontology Lookup Service [EB/OL]. [2022-01-17]. https://www.ebi.ac.uk/ols/index.

[11] WHEELER D L, BARRETT T, BENSON D A, et al. Database resources of the National Center for Biotechnology Information [J]. Nucleic Acids Research, 2006, 34(D1): D173-D180.

[12] VALENTIN F, SQUIZZATO S, GOUJON M, et al. Fast and efficient searching of biological data resources-using EB-eye [J]. Brief Bioinform, 2010, 11(4): 375-384.

[13] SQUIZZATO S, PARK Y M, BUSO N, et al. The EBI Search engine: providing search and retrieval functionality for biological data from EMBL-EBI [J]. Nucleic Acids Research, 2015, 43(W1): W585-W588.

[14] MADEIRA F, PARK Y M, LEE J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019 [J]. Nucleic Acids Research, 2019, 47(W1): W636-W641.

[15] BAXEVANIS A D. Searching NCBI Databases Using Entrez [J/OL]. Curr Protoc Hum Genet [2021-12-03]. https://doi.org/10.1002/0471142905.hg0610s51.

[16] EMBL EBI impact report 2021 [EB/OL]. [2021-12-03]. https://www.embl.org/documents/wp-content/uploads/2021/10/EMBL-EBI-impact-report-2021.pdf.

[17] LANDRUM M J, CHITIPIRALLA S, BROWN G R, et al. ClinVar: improvements to accessing data [J]. Nucleic Acids Research, 2020, 48(D1): D835-D844.

[18] TUNYASUVUNAKOOL K, ADLER J, WU Z, et al. Highly accurate protein structure prediction for the human proteome [J]. Nature, 2021, 596: 590-596.

[19] VARADI M, ANYANGO S, DESHPANDE M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models [J]. Nucleic Acids Research, 2022, 50(D1):

D439-D444.

［20］UNESCO Recommendation on Open Science［EB/OL］.［2021-12-03］. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en.

作者简介

吴思竹，女，1981年生，博士，研究馆员，研究方向：医学科学数据管理和应用，E-mail：wu.sizhu@mail.imicams.ac.cn。
王安然，女，1993年生，硕士，助理研究员，研究方向：医学科学数据标准化。
修晓蕾，女，1993年生，硕士，研究实习员，研究方向：医学知识图谱建设。
钱庆，男，1970年生，硕士，研究员，研究方向：医学科学数据管理、医学知识组织体系建设。
周伟，男，1970年生，硕士，高级工程师，研究方向：人口健康科学数据管理。

Construction and Implications of Biomedical Scientific Data Centers in Europe and the United States

WU SiZhu[1] WANG AnRan[1] XIU XiaoLei[1] QIAN Qing[1] ZHOU Wei[2]
(1. Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, P. R. China; 2. National Population Health Data Center, Beijing 100005, P. R. China)

Abstract: This paper comprehensively analyzed the construction and development experience of the National Center for Biotechnology Information of the United States and the European Bioinformatics Institute, which have important international influence, in the construction and development of biomedical data resources, tool utilization and sharing services. It summarized enlightenment and suggestions for the construction of biomedical science data centers in China from six aspects, such as funding and personnel support, IT infrastructure, resource construction, tool research and development, standards and norms, and user service.

Keywords: Biomedical Science; Scientific Data Center; Data Management; Data Sharing

（收稿日期：2022-04-12）

书 讯

《汉语主题词表》

《汉语主题词表》自1980年问世以后，经1991年进行自然科学版修订，在我国图书情报界发挥了应有作用，曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要，由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，自2009年开始进行重新编制工作，拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表（工程技术卷）》共收录优选词19.6万条，非优选词16.4万条，等同率0.84，在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表（自然科学卷）》共收录专业术语12.4万条，包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域，收词系统、完整，语义关系丰富、严谨，每条词汇都有相应的学科分类号表现其专业属性，并与同义英文术语对应。同时，建立《汉语主题词表》网络服务系统，提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘，是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表（工程技术卷）》已于2014年由科学技术文献出版社出版，分为13个分册，总定价3 880元。《汉语主题词表（自然科学卷）》已于2018年5月由科学技术文献出版社出版，分为5个分册，总定价1 247元。两卷均可分册购买。