DOI: 10.13376/j.cbls/2023168
文章编号：1004-0374(2023)12-1545-08

李亦学，研究员，博士生导师。现为广州实验室研究员、国家生物数据中心体系粤港澳节点平台首席科学家兼主任，中国生物信息学会(筹)副理事长，上海生物信息学会理事长，上海交通大学教授，复旦大学遗传学教育部协同创新中心前沿生物技术部主任。主要研究领域为生物信息学、基因组学、精准医学、人工智能、大数据、知识图谱等。曾任国家“十五”863计划生物和农业技术领域生物信息技术主题专家组组长，国家“十一五”863计划生物医药技术领域专家组专家；国家蛋白质科学研究重大专项“模式生物和细胞等功能系统的系统生物学研究”“代谢生理活动与病理过程中信号转导网络的系统生物学研究”两任专项项目首席科学家。获得上海市自然科学奖一等奖、二等奖，教育部自然科学奖等，并荣获全国五一国际劳动奖章，上海市劳动模范，第一批上海市“科教兴市领军人才”等。

生物医学大数据生产要素价值的实现：从数据元素起步

张国庆[1]，赵国屏[1]，李亦学[1,2]*
(1 中国科学院上海营养与健康研究所，生物医学大数据中心，上海 200031；2 广州实验室，广州 510005)

摘　要：基因组学 / 系统生物医学、转化医学、精准医学时代以来形成的生物医学大数据不仅是生物医学领域开展数据密集型研究的基石，成为与人口健康、社会发展和国家安全相关的战略资源，而且还是利用人工智能赋能“大健康”产业发展的核心生产要素(常简称为“数据要素”)。生物医学数据元素具有与生物和医学相关的“跨尺度、多源性、高维度、细粒度”等异质性复杂体系特征，因此，具有4V特征(Volume、Velocity、Variety、Veracity)的海量生物医学数据的数据元素必须经标准化规范整合并供共享分析，才能将海量生物医学数据质变转化为生物医学大数据，发挥生产要素的功能，实现生产要素的价值。这个价值释放的“要素化”过程，面临着特有的机遇与挑战，特别是已经成为生物学与健康医疗大数据最核心的基础的多组学及多模态数据，与欧美相比，我国数据“多而不强”，由于开放共享程度低、集中程度不高，难以评估数据质量。数据库是生物医学数据共享的主要载体，其数据来源和共享模式直接影响数据要素的价值释放过程。数据中心是数据库的建设及运维主体，也是各类数据元素转换为适用各类应用场景的数据要素的重要参与者和推动者，处于数据要素化不可或缺的核心环节。在从数据元素转换到数据要素的过程中，我们面临着存量数据规模与数据规范化集成的治理能力不匹配、已开放的数据规模与数据分析挖掘的治理能力不匹配的挑战，需要在数据、数据库、数据中心三个层面上加强数据治理和数据共享等基础性工作。我们建设了1(套 整合交互共享导向的数据资源服务体系)-2(个 标准化数据分析平台)-3(种 科学 / 技术问题驱动的健康医学数据治理平台)-X(类 面向应用场景的智能分析服务体系)的生物医学大数据技术体系，秉承“安全管理、信息共享、标准增值、技术创新、尊重产权、高效利用”理念，努力将数据中心从成本中心转换为价值中心，可为生物医学大数据“要素化”提供借鉴。

关键词：生物医学大数据；数据元素；数据要素；数据共享；数据治理
中图分类号：Q-3；R-05　　**文献标志码：**A

收稿日期：2023-11-29；**修回日期：**2023-12-15
基金项目：国家重点研发计划(2021YFF0703802)
***通信作者：**E-mail: yxli@sibs.ac.cn

Value realization of biomedicine big data as the production factor: starting from data elements

ZHANG Guo-Qing[1], ZHAO Guo-Ping[1], LI Yi-Xue[1,2]*

(1 Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China; 2 Guangzhou Laboratory, Guangzhou 510005, China)

Abstract: Biomedical big data (BMBD), which has emerged since the inception of genomics, systems biology, translational medicine, and precision medicine, not only forms the foundation of data-intensive research in the biomedical field, but also serves as a strategic asset with far-reaching implications for public health, societal progress, and national security. Moreover, it plays a pivotal role in driving the expansion of the "big health" industry through the application of artificial intelligence. Biomedical data exhibit heterogeneous complexity characterized by their relevance to biology and medicine across different scales, multiple sources, high dimensions, and fine granularity. Therefore, massive biomedical data with the 4V characteristics (Volume, Velocity, Variety, Veracity) must undergo standardization, specification, integration, and sharing for collaborative analysis. This process is essential to transform the sheer volume of biomedical data into biomedical big data, enabling these data to function as essential production factors and realize their intrinsic value in the realm of healthcare and medical research. The process of transforming massive data into BMBD with the value as production factors presents distinctive opportunities and challenges, especially in the context of multi-omics and multimodal data, which now serve as the foundation of biology and healthcare big data. In comparison to Europe and the United States, China's data landscape is characterized as “abundant but not robust,” marked by low levels of openness and lack of integration, which makes interactive data search and easy data assess a challenging task. Databases play a pivotal role as the primary conduits for sharing biomedical data, and their data sources and sharing models directly influence the process of extracting value from data elements. Data centers are responsible for both constructing and maintaining databases, playing a pivotal role as significant contributors and advocates in the transformation of diverse data elements into adaptable key production factors suitable for various application scenarios. As we strive to convert data elements into production factors, we face challenges arising from disparities between governance capacity and the scale of existing data and data standardization / standardized data collection, as well as a mismatch between governance capacity and the scale of open data and data analysis and mining. It is imperative to strengthen the foundational aspects of data governance and data sharing across three levels: data, databases, and data centers. We have implemented a 1 (set of integrated interactive sharing guided data resource system)-2 (sets of standardized data analysis platform)-3 (sets of science and technology driven governance platform for health medicine big data)-X (sets of application scenario based knowledge mining AI analysis systems) BMBD technological engineering system, guided by the principles of “safety and security assured management, information sharing, value-added standardization, technological innovation, IP respected, and efficient utilization”. Our ongoing efforts are directed towards transitioning the data center from its currently cost-burdened status to a value-added level, poised to offer invaluable insights for unlocking the value of BMBD.

Key words: biomedical big data (BMBD); data elements; data factors; data sharing; data governance

基因组学 / 系统生物医学 (genomics/systems biomedicine)、转化医学 (translational medicine)、精准医学 (precision medicine) 时代以来形成的生物医学大数据 (Biomedicine Big Data) 来自于生物学 (生命科学)/ 生态学 (环境科学)、医学和药学以及环境和社会科学等极其宽广范畴的科学研究、技术、临床与产品研发以及真实世界数据；仅仅在科学研究方面，就包含了生命科学研究、基础医学研究、临床医学研究、预防医学 / 流行病学研究等多个方面的数据。随着数据采集与利用技术的进步，生物医学数据的拓展，既趋向整合从个体到群体的研究型数据 (research data)，还趋向整合从部分到全生命周期的真实世界 / 真实生活数据 (real-world/real-life data)。因此，生物医学大数据不仅对今天的基础与应用研究越来越重要，是生物医学领域开展数据密集型研究的基石，成为与人口健康、社会发展和国

家安全相关的战略资源 (strategic resource)，而且还是利用人工智能赋能“大健康”产业发展的核心生产要素 (key production factors)，亦可简称为数据要素 (data factor)。

当数据作为新型生产要素时，不像土地、劳动力资本等传统生产要素可以表现为具体的实物个体价值，需要集成整合成为信息，乃至分析挖掘形成知识才能发挥价值。在数据集成整合、共享传播过程中涉及的众多主体对价值都有或多或少的贡献，而且由于其独特的低成本复制性，导致数据要素具备依附倍增性 [1]，例如“互联网 +”“AI for”等近年来国家大力引导的方向，都可以视为是数据与其他生产要素结合后能够释放更大的价值；而且数据要素存在三次价值释放过程，第一次是“数据支持业务贯通”，第二次是“数据推动数智决策”，第三次是“数据流通对外赋能”[2]。这种一般性的数据向实现生产要素价值转化过程的特征，在生物医学大数据方面也同样可以看到。但是，更为重要的是看到其特性，并在此认识的基础上，实现数据的“集成整合成为信息，乃至分析挖掘形成知识，并形成价值”的生产要素转化，也就是所谓的“Data-Information-Knowledge-Wisdom”的 DIKW 过程。

今天意义上的生物医学海量数据不仅具有“量大、实时、质杂、有偏”等大数据 4V 特征 [3]，还有其他类型大数据难以企及的“跨尺度、多源性、高维度、细粒度”等异质性复杂体系特征 [4]。因此，在将其转化成为生产要素时，除了要完善数据确权、流通交易、收益分配等数据基础制度体系之外，首先需要深入讨论“数据元素”(data element) 的特点及其在数据向生产要素转化过程中面临的技术与工程需求，才能推动海量生物医学数据到生物医学大数据的质的转变，切实解决数据供给的源头问题，发挥生产要素的功能，实现生产要素的价值。为了叙述简便，本文将上述从“海量生物医学数据”的数据元素向“生物医学大数据”要素转化的质变过程，称之为“生物医学大数据要素化”。将首先剖析转化中的基本特征，以及作为数据元素主要载体的数据库及其共享方式的特点，处于数据要素化核心环节的国内外数据中心的差异，并阐述我国数据要素价值释放所面临的挑战，最后分享我们近六七年来探索建立生物医学大数据治理体系的实践经验。

1　生物与健康医疗数据元素与生物医学大数据要素化

生物医学数据，从数据科学角度看，最基本的单元是数据元素 (data element，也称为数据元)，是用一组属性描述其定义、标识、表示和允许值的数据单元，在一定语境下，通常用于构建一个语义正确、独立且无歧义的特定概念语义的信息单元。只有将这些元素按照标准 (standards) 规范化集成，成为具有信息内涵的可供分析挖掘形成知识的“大数据”，同时以规范的方式开放共享，才能在科研及产业发展中释放价值，也就是我们所谓的大数据的要素化。我们在本文内把生物医学数据简化分为生物和健康医疗两类数据元素，分别阐述其特点。

由于历史原因的影响，目前普遍强调的生物大数据主要是指测序相关的多组学数据。INSDC 是开放数据领域最著名的全球倡议之一，在核酸序列及测序数据共享方面发挥了巨大作用 [5]。百慕大原则是人类基因组计划中大规模施行的数据共享原则，虽然没有达到“每日共享数据”的目标，但是其倡导的数据共享对质量控制乃至总体目标的达成所起到的推动作用，仍然对现在的大型科研项目具有参考价值 [6]。此外，ProteomeXchange 从蛋白质组的角度推动了数据开放共享 [7]，并且与 INSDC 的成员库一样，被众多出版集团推荐为论文伴随数据的数据仓库。在国际数据协会 / 联盟、期刊以及大型国际合作项目的影响下，形成了以美国 NCBI 和欧洲 EMBL-EBI 为核心的存储共享机制，对测序、蛋白质组以外的生物数据共享具有很强的示范效应。本世纪以来，我国产出的组学数据日益增多，但由于各种原因，研究数据在公开发表时，基本都会按照期刊对数据开放共享的要求，递交存储于国外数据库；相对而言，未公开发表的数据、分析结果可能由研究团队或检测机构碎片式保存，原始数据甚至可能仅由检测机构保存，在项目结题时会按照科技部等经费主管部门的要求汇交到国家科学数据中心。

健康医疗数据主要包括队列研究数据、临床研究数据，以及以临床诊疗记录为主的真实世界数据，不仅强调个体在疾病、健康状态的特征及表型，而且近年来也不断采集影像及分子表型。以 Framingham 研究为例，该队列已有 75 年历史，先后引入 CT 和 MRI 等影像技术，以及 SNP 检测、甲基化检测和全基因组检测等组学检测技术 [8]。20

世纪 90 年代以来，以国家行为部署的队列研究迅速增长，例如英国的 UK Biobank (2006 年，50 万人，https://www.ukbiobank.ac.uk/) 和 Our Future Health (2023 年，500 万人，https://ourfuturehealth.org.uk/)，美国的 All of Us (2016 年，100 万人，https://allofus.nih.gov/)，以及日本、法国、瑞典、挪威等国的队列 (均在 50 万人左右)。我国近年来先后启动了多个大型队列，例如复旦泰州自然人群队列 (20 万人)、中国科学院精准健康队列 (100 万人) 等。我国还有部分与国际团队共同开展的队列研究，如北大与牛津大学共建的 CKB 慢性病队列 (50 万人) 等。这些“新时代”建设的队列，大多引入了多组学或多模态数据，采用先采集样本，再按需组学检测的方式持续获取数据。相对于流行病学调查问卷及实验室检查等数据，组学及影像等数据具备文件巨大、信息丰富的特点，因此在队列等健康医疗研究中常常单独管理多组学及多模态数据，甚至直接存放在第三方数据中心。这些健康疾病相关的数据，由于包含大量可能涉及个人信息的表型，很少采用与生物数据相似的开放共享方式，大多以协议的方式进行受控共享。

由此可以看到，虽然国内数据产出规模大，但是由于开放共享程度低、集中程度不高，难以评估数据质量，因此无论是生物大数据，还是健康医疗大数据，在转换为数据要素时，大而不强，价值释放过程存在一定的困难。

2 生物医学数据库的数据共享模式

数据库是生物医学数据共享的主要载体，其数据来源和共享模式直接影响数据要素的价值释放过程。

目前提供共享服务的数据库所收录的数据按照来源可以分为汇交型和项目门户型两大类。美国 NCBI、欧洲 EBI 等数据中心经过数十年的发展，超大规模的数据持续汇交到他们运行管理的数十个生命组学及大 / 小分子相关的数据仓库，并通过基础分析工具提供相应的搜索与计算服务，得到了国际期刊及学者的普遍认可。另一方面，英国 UK BioBank (50 万人的健康调查及多模态数据) 和美国 TCGA (万名肿瘤患者的多组学数据) 等数据库，依托于大型科研项目产出的高质量的规范化采集的数据，面向全球科研人员提供数据共享申请服务，一旦获批后研究者就可以基于共同签署的协议来开展研究，这种做法使得数据库享有广泛的国际影响力，形成了巨大的科学和社会效益。

数据库主要包括完全公开和受控共享 2 种开放共享方式，而这与其收录的数据涉及的政策法规及知识产权 (含研究者利益保护) 密切相关。首先，在生物大数据中，人类遗传相关的数据需要遵循各国的相关法律法规，例如美国 HIPPA[9]、欧盟 GPDR[10]、《中国人类遗传资源管理条例》[11-12] 等政策法规，传染性疾病及高生物安全等级相关的病原体数据需要遵循生物安全法，除这些数据需要受控共享外，其他类型的数据开放共享状态更多的是取决于知识产权方面的考虑，倾向于完全公开。而在健康医疗数据中，涉及到大量个人信息、敏感信息，同样需要遵循相关法律法规，在开放前需要进行脱敏及匿名化处理，而且大多以受限的方式进行共享。数据安全问题是受限开放的数据库必须面临的重要问题，而且随着数据应用能力的提升，社会各界对于数据安全的担忧，可能会促使数据共享越来越倾向于受控共享的方式。以人类遗传资源为典型，数据共享需要由数据提交者、管理者或者监管部门进行评估、审核并签订协议后才能实施，在这个过程中如何通过技术手段来评估数据安全风险并辅助共享决策，将是未来的挑战之一。另一方面，“完全公开”的数据共享，还需要结合数据库的政策来判定，例如 GISAID (https://gisaid.org/) 要求在使用数据时要致谢所有数据贡献者，而 NCBI SRA (https://ncbi.nlm.nih.gov/sra) 中可能包含了极少数有隐性共享条件的数据 (例如，要求在发表基于特定数据的研究成果前与提交者联系)，而这要求用户在共享得到的数据的时候要尊重数据提交者、数据库等相关方的约定或要求，特别是在涉及论文尚未发表但是数据已经公开的数据时，应格外小心。

基于数据库的开放共享是数据价值释放的有效手段，这个过程伴随的数据安全、知识产权、法律法规等问题，既需要数据治理以及数据共享等技术的升级，也需要制度保障。也唯有这样，才能更好地体现数据价值。

3 生物医学数据中心建设趋势

数据中心是数据库的建设及运维主体，也是各类数据元素转换为生产要素的重要参与者和推动者，处于数据要素化不可或缺的核心环节。

美国 NCBI (https://ncbi.nlm.nih.gov/)、欧洲 EMBL-EBI (https://www.ebi.ac.uk/)、日本 NIG DDBJ (https://www.ddbj.nig.ac.jp/) 作为 INSDC 的首批成员，各自

运维了数十个国际性的数据库，尤其是 NCBI 和 EMBL-EBI，长期面向全球提供递交、下载、分析等开放数据服务，以及数据科学技术的研究，网络 / 会议 / 培训 / 新媒体等多途径的传播推广服务，在生物医学领域拥有广泛的影响力。美国 NCI TCGA (https://www.cancer.gov/ccg/access-data)、欧洲 Tara Oceans[13] 等大型科学项目的数据类型不仅局限在单一组学数据，甚至不仅仅是多组学生命组学数据，这对传统的数据中心提出了新的要求，数据中心从原先的集中式管理，正在呈现多学科共建、多中心共运营的趋势，如前文所述，其组学数据存放在 NCBI 和 EBI 的相关数据库中，表型数据自建或者存放在其他的数据库中，这样可以针对不同数据的特点，选择更合适的开放时间和共享政策。英国由 72 个生物学、医学和数据工程师协会等机构共同参与建设的 HDRUK (https://www.hdruk.ac.uk/)，整合临床医疗健康数据 (即英国全民医保 NHS 数据) 和多组学等生物医学科研数据 (主要是 MRC 和 EBI 数据)，利用“英国工程与自然科学研究理事会 (EPSRC)”的数据信息技术，以分布于全国各地的各具特色的 9 个数据研究中枢为核心，建立起了英国健康数据的基础治理体系和科研服务设施，为提升生物医学研究水平和临床医疗水平，改善人民健康，发挥了重要的作用。这一体系的建设，主要依靠在原先各自独立的数据中心之间建立标准化的、互通互联的合作关系。譬如，众多大型项目的多组学数据同样是存放在 NCBI/EBI 等数据中心的，而非另起炉灶式地搞新的大型基建或设立新的大型项目，从而大大提高了人力与财力的利用效率。这一体系的另一个重要的基本特点 (与其他国际医学数据中心类似)，就是与生物数据中心密切合作，而且向所有符合要求的用户开放。

我国于“九五”期间开始开展人类基因组研究时，就有专家倡议建设生物医学数据中心 (当时基本称之为“生物信息中心”)[14]，并最早支持北京大学建立与 EBI 相连的节点 [15]。此后，一方面地方与研究机构合作建立了若干生物信息中心，如上海市科委与中国科学院上海生命科学研究院合作建立的上海生物信息技术研究中心，另一方面，科技部通过平台支撑项目支持成立了基因组、微生物和人口健康等生物医学相关的国家科学数据中心，支撑科技部的项目数据汇交工作。当前阶段，这些生物数据中心的职责以存储数据为主，努力与国际接轨，已具备一定的数据开放共享服务的能力和影响力。

在健康医疗数据方面，国家卫计委 / 卫健委 2016 年提出了“一个国家数据中心，七个区域中心，并结合各地实际情况，建设若干个应用发展中心”的建设任务，也就是实施“1+7+X”健康医疗大数据应用发展的总体规划。此外，在科技部及国家卫健委布局国家临床医学研究中心中，部分中心建立了相应的医学数据中心，如国家神经系统疾病临床医学研究中心等已经提供了在线数据共享系统 (https://www.ncrcnd.org.cn/)；上海申康医院发展中心也资助了数十个专病数据库 (含样本库)。与国外通过在线系统进行数据共享相比，国内医学数据中心的数据开放政策和方式明显不同。这些已建或在建的临床医学及流行病学研究相关的数据中心，绝大多数只向机构内部或特定用户开放，缺少与生物数据中心的资源共享与合作交流，不能有效地与临床数据互联互通，导致数据完整度和可信度不足以直接支撑医院及医生开展临床研究、循证医学研究和“真实世界数据研究”，仍然需要回溯至医院数据中心获取更原始的数据。

如上所述，国家、地方及部分科研和医学机构组建了各级各类生物学或医学大数据中心，已经汇集了一定甚至相当规模的数据资源。但是，这些数据中心目前仍然处于“数据汇交汇聚”的阶段，绝大多数还处于从“碎片化”到“孤岛化”的阶段，极少部分可达到“烟囱化”的高度，需要进一步以应用为导向，加强数据标准化和质量控制等基础性工作，突破数据多样性和整合程度受限严重、数据开放共享范围和用户群体有限的瓶颈，在数据要素化过程中发挥中坚力量。

4　生物医学数据要素化的挑战与机遇

数据要素有三次价值释放过程。以医疗数据为例，首先医疗数据在医院内部流转并支撑相关业务，完成数据要素流通的基础工作，通过支持业务贯通第一次释放价值；然后通过统计分析、知识图谱、人工智能等技术手段，以智慧医院、临床 (辅助) 决策系统等为载体，辅助医院及医务工作者决策和管理，完成第二次价值释放；最后，由于数据特有的低成本复制特性，通过多中心研究、区域医联体等多种形式的开放共享，医疗数据流通对外赋能，第三次释放价值 [2]。长期以来，除医疗以及基因检测等数据能够直接释放第一次价值外，其他的生物医学数据更多的是通过科学发现来体现价值，很难像其他生产要素一样，可以以定价的方式直接体现

其价值；但是，得益于生物医学大数据领域长期以来的开放共享理念，以及生物信息学、计算生物学、人工智能等技术手段的支持，如果能够供给高质量的数据，其第二次、第三次价值释放反而更易达成。

针对我国生物医学大数据的现状，我们认为在生物医学数据要素化并进一步释放价值时，存在两类主要挑战：

(1) 存量数据规模与数据规范化集成的治理能力不匹配，存在大量的治理不充分的质量参差不齐的数据；特别是，我国的生物和健康医疗两类数据中心普遍缺乏高效技术手段和相应的协同体系来进行数据集成，提供标准统一、收集规范的高质量生物医学大数据。典型现象为：数据中心实际拥有大量数据，但是需要大量人力物力和时间才能转化为可取用的“大数据”。

(2) 已开放的数据规模与数据分析挖掘的治理能力不匹配，一方面有限的可接触到数据的分析团队不足以充分挖掘数据价值，另一方面国内众多难以获取数据的分析团队只能选择国际数据进行研究。典型现象为：我国学者利用 TCGA、UK Biobank 及其他国外数据开展了大量研究，其成果数量和质量远高于利用我国数据开展的研究。

上述挑战严重制约了我国生物医学数据资源的生产要素能力，形成大量的数据烟囱和孤岛。最近数十年国内外经验教训已经证明，生命科学与医学大数据需要深度融合，尤其是挖掘人及模式动物的多组学、多模态数据，并且需要生物、医学、工程、计算等多学科的协作，联通数据烟囱及孤岛，才能加速生物医学数据要素的形成和供给。具体而言，建议以应用为导向，在数据元素、数据库、数据中心等三个层面上加强数据治理和共享工作：

(1) 数据治理技术：在数据层面上应对数据进行规范化采集和治理，形成高质量的数据；在数据库层面，项目门户型数据库，尤其是汇交型数据库，应将数据治理技术前置，把异源异构的数据治理为分层分级的可共享的数据；在数据中心层面，应建立数据治理技术所需要的存储、计算、服务与安全等配套基础设施，建立数据治理及质量评价标准，针对质量不高或者难以判定的数据制定合理的管理策略。

(2) 数据共享技术：在数据层面上，去芜存菁，将高质量的数据合法合规地开放共享，增加可要素化的数据元素；在数据库方面，进一步完善数据共享方式和途径，对高价值数据强化可交互取用的实时共享方式；在数据中心层面上，应该在基础设施、技术研发、评价体系等各方面强化应用导向，鼓励并引导多层次、多尺度的数据共享，将“数据中心”通过要素化，从“成本中心”向“价值中心”转化。

5 实践与思考

经过多年建设，我们已经建成了 1-2-3-X 的生物医学大数据技术体系 (图 1)。“1”是生物数据汇交技术体系，我们在 2016 年启动了组学数据百科全书 NODE (https://www.biosino.org/node/) 的建设，2016 年 4 月首个组学数据上线，截止到 2023 年 10 月底，NODE 数据规模已经达到 4 PB，访问人次达到 5 000 万；此外，在 2019 年上线了微生物系统分类与基因组数据库 eLMSG (https://www.biosino.org/elmsg/)，2020 年上线了合成生物学元件数据库 RDBSB (https://www.biosino.org/rdbsb/)，这 3 个数据库有效支撑了 *Cell*、*Nature*、*Science* 等知名生物医学期刊论文数据的伴随数据发表，由此，我们实现了从生命多组学 (原始数据)-(拼接后的) 基因组 -(分析挖掘后的) 生物元件的生物数据汇交技术体系。为了提升数据规模化和工程化的分析挖掘能力，我们研发了 2 个生物信息分析云平台，分别是精准医学交互式分析云平台 ViPMAP 和微生物生物信息分析云平台 iMAC，整合了超过 200 项软件，形成了近 50 条各类组学数据的分析流程，通过高性能计算环境提供免费服务，并已与国产云计算平台进行适配和部署。与生物大数据强调开放共享不同的是，健康疾病数据的重点是数据集成与治理，为此，我们先后与合作单位联合研发了面向深度表型测量与分子表型提取的人类表型组数据采集与共享平台，面向现场规范化采集的自然人群队列数据采集与管理平台，致力于多来源异构数据集成与治理的专病数据库 / 队列数据集成与共享平台，由此形成了 3 种面向不同应用场景的大数据技术平台，并且已经与多家科研单位及医疗机构的已有信息系统对接，包括 REDCap[16] 等开放临床研究数据管理系统、商业数据管理系统或者定制的数据库，通过在合作单位的部署，形成数据采集、管理、共享、应用的完整的健康医疗大数据解决方案。

通过 1 套整合交互共享导向的数据资源服务体系，2 个标准化数据分析平台，3 种科学 / 技术问题驱动的健康医学数据治理平台的研发与运行维护，我们具备了以生物信息分析为主要手段的生物数据治理能力，以机器学习、人工智能及大模型为主要

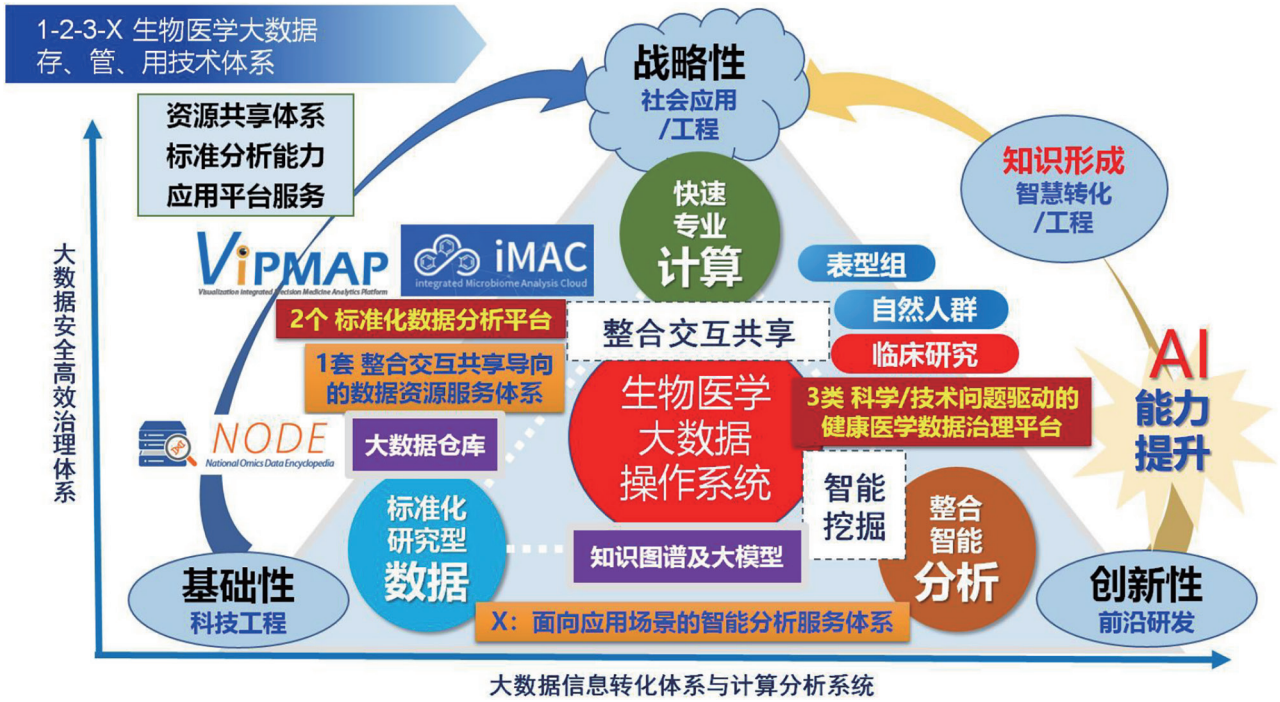

图1　1-2-3-X生物医学大数据存、管、用技术体系

手段的健康医疗数据治理能力，从而研发了（多）X个的生物医学数据库和知识库，包括SMDB、PGG.Han/PGG.SV、Transcirc、dbDEMC等，并通过https://www.biosino.org/为相关的应用场景提供开放服务，多个数据库已经成为国家基因组科学数据中心的成员库[17]。

在1-2-3-X的生物医学大数据技术体系建设过程中，我们通过数据治理和共享，把原始数据转换为更易发挥价值的知识，并且以技术合作的方式应用部署，秉承“安全管理、信息共享、标准增值、技术创新、尊重产权、高效利用”理念，在数据资源、技术、管理等多方面开展了尝试，已经初步实现了我们倡议的“以递交为基础、以整合为导向的数据存储中心，以主题为基础、以交互为导向的数据共享中心，以及以传统信息技术为基础、以前沿信息技术为导向的下一代生命科学数据转化中心”[18]。这些工作在供给高质量数据方面，可为生物医学大数据“生产要素”化提供借鉴。

最后，数据中心作为生物医学大数据的管理和实施单位，其建设、运行、资助和评价模式的差异，将极大地影响数据生产要素的数量和质量。国内外医学数据中心的运行模式（即生物学数据与医学数据的交汇共通）和开放共享（即开放的范围以及共享使用的便利程度）存在根本性的差异，而且由于资助模式（经费支持是否长期、稳定、足额）及评价体系（主管/监管/资助机构评价与服务对象评价）的影响，我国生物及医学数据中心更应以“应用导向”，利用人工智能大模型等新一代智能计算技术，加强了我们对数据的治理与应用能力，推动“数据密集型研究范式”的形成，实现“数据驱动的知识发现”的愿景，施惠于人民的大健康事业。

[参　考　文　献]

[1] 任保平, 李婧瑜. 数据成为新生产要素的政治经济学阐释. 当代经济研究, 2023, 339(11): 5-17

[2] 王泽宇, 吕艾临, 闫树. 数据要素形成与价值释放规律研究. 大数据, 2023, 9: 33-45

[3] 潘逸航, 郑子龙, 张国庆, 等. 临床研究大数据治理平台的建设与实践. 中国卫生信息管理杂志, 2022, 19: 918-24

[4] 赵国屏, 李亦学, 陈大明, 等. 生物医学大数据的态势与展望[M]//中国科研信息化蓝皮书 2020. 北京: 电子工业出版社, 2020: 11-30

[5] Cochrane G, Karsch-Mizrachi I, Nakamura Y, et al. The international nucleotide sequence database collaboration. Nucleic Acids Res, 2011, 39: D15-8

[6] Maxson Jones K, Ankeny RA, Cook-Deegan R. The Bermuda Triangle: the pragmatics, policies, and principles for data sharing in the history of the human genome project. J Hist Biol, 2018, 51: 693-805

[7] Vizcaino JA, Deutsch EW, Wang R, et al. ProteomeXchange provides globally coordinated proteomics data submission

and dissemination. Nat Biotechnol, 2014, 32: 223-6
[8] Andersson C, Johnson AD, Benjamin EJ, et al. 70-year legacy of the Framingham Heart Study. Nat Rev Cardiol, 2019, 16: 687-98
[9] Health Insurance Portability and Accountability Act of 1996 (HIPAA)[EB/OL]. https://www.cdc.gov/phlp/publications/topic/hipaa.html
[10] General Data Protection Regulation[EB/OL]. https://gdpr-info.eu/
[11] 中华人民共和国人类遗传资源管理条例[EB/OL]. https://www.gov.cn/zhengce/content/2019-06/10/content_5398829.htm
[12] 人类遗传资源管理条例实施细则[EB/OL]. https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/bmgz/202306/t20230601_186416.html
[13] Sunagawa S, Acinas SG, Bork P, et al. Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol, 2020, 18: 428-45
[14] 郝柏林. 建议尽快组建国家级生物医学信息中心. 中国科学院院刊, 2000, 15: 133-4
[15] 罗静初. 顾孝诚教授与北京大学生物信息中心. 中国科学:生命科学, 2022, 52: 1555-60
[16] Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform, 2009, 42: 377-81
[17] CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2023. Nucleic Acids Res, 2023, 51: D18-28
[18] 张国庆, 李亦学, 王泽峰, 等. 生物医学大数据发展的新挑战与趋势. 中国科学院院刊, 2018, 33: 853-60