

国外人口健康领域科学数据共享平台建设特点及启示

付磊¹ 李金斌¹ 王戎² 尚小溥³ 尹岭¹

(1. 中国人民解放军总医院, 北京 100853; 2. 国家人口计生委科学技术研究所, 北京 100081;
3. 北京交通大学, 北京 100044)

摘要: 以承载生物信息、临床试验、循证医学三类数据并实施共享的典型数据平台NCBI、CT、UptoDate为对象, 从外部政策环境、数据组织管理、数据采集更新机制、数据共享利用等方面, 总结国外人口健康领域科学数据共享平台的建设特点, 并在分析国内现状和不足的基础上, 对我国人口健康领域科学数据共享平台的建设提出建议。

关键词: 人口健康; 医药卫生; 科学数据; 数据共享; 平台建设

中图分类号: R589

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2017.05.012

Constructing Characteristics About Scientific Data Sharing Platform in the Field of Population Health in Foreign Countries and Its Inspiration to Us

FU Lei¹, LI Jinbin¹, WANG Rong², SHANG Xiaopu³, YIN Ling¹

(1. Chinese PLA general hospital, Beijing 100853; 2. National Research Institute for Family Planning, Beijing 100081; 3. Beijing Jiaotong University, Beijing 100044)

Abstract: This paper takes NCBI, CT and UptoDate which implements the sharing of biological information, clinical trials and evidence-based medical data as the typical objects, summarized the characteristics of scientific data sharing platform in the field of population health abroad from external policy environment, data reorganization and management, data collection and update, data sharing and utilization, and Then, based on the analysis of the current situation and shortcomings in China, put forward some suggestions to the construction of the scientific data sharing platform in the field of population health in China.

Keywords: population health, medical and health, scientific data, data sharing, platform construction

1 引言

科学数据指在调查、实验、探测等科技活动中或通过其他方式所获取的反映客观世界的本质、特征、变化规律等原始数据, 以及根据不同

科技活动需要进行系统加工整理的各类数据集^[1]。科学已步入了“大数据”时代, 科学数据是大数据时代最基本、最活跃、影响最广泛的科技创新资源, 具有重要的科学、经济和社会价值, 通过海量科学数据的交换、整合、分析, 新的知识和

作者简介: 付磊 (1982—), 男, 博士, 中国人民解放军总医院博士后, 主要研究方向: 医学信息学; 李金斌 (1989—), 男, 硕士, 中国人民解放军总医院工程师, 主要研究方向: 医学信息学; 王戎 (1978—), 女, 博士, 国家人口计生委科技学术研究所助理研究员, 主要研究方向: 医学信息学; 尚小溥 (1984—), 男, 博士, 北京交通大学经济管理学院讲师, 主要研究方向: 管理科学与工程; 尹岭 (1955—), 男, 博士, 中国人民解放军总医院主任医师, 教授, 主要研究方向: 医学信息学、神经内科学 (通讯作者)。

基金项目: 科技基础性工作专项重点项目“科技基础性工作数据资料集成与规范化整编” (2013FY110900); 国家人口与健康科学数据共享平台专项课题“人口健康平台临床数据资源建设发展策略研究” (2016NCMIZX07)。

收稿时间: 2017年7月14日。

规律被不断发现,新的意义和价值被不断产生和创造。作为最活跃的科学研究领域之一,人口健康领域的科学数据被广泛应用于药物研发、疫情监测、公共健康监控、临床实验数据分析、药械安全性与有效性以及卫生经济学评价等多方面。作为科学数据收集、组织、存储、加工、传播和利用的关键基础设施,科学数据共享平台的建设受到越来越多的重视。追踪并研究世界范围内人口健康领域科学数据共享平台建设情况和新进展,学习引进新思路、新技术、新方法,有利于大幅提高我国在该领域的建设水平,从而少走弯路,实现跨越式发展。

2 调查对象与方法

人口健康领域科学数据类型复杂,形式多样。随着人类对自身研究的不断深入,以及现代计算机技术、网络技术、信息技术的不断发展,各种机构和组织建立的用于收集、存储、管理人口健康数据的各类型平台日渐增多。然而,很多平台仅局限于各自机构内部使用,外部交流与共享的范围和内容都十分有限。目前,人口健康领域科学数据共享实现较好的主要包括生物信息类数据、临床试验类数据和循证医学类数据。本文以承载这三类数据的典型共享平台NCBI (National Center of Biotechnology Information)、CT (Clinical Trial) 和UptoDate为主要调查对象。在调查方法上,以网络调查法访问各平台为主,辅以文献调研。由于关注的学科领域不同,各平台的数据内容、数据规模差异较大,相关内容并不作为调查重点,而主要从外部政策环境、数据组织管理、数据采集更新机制、数据共享利用等方面,分析总结相关特点,为我国人口健康领域科学数据共享平台的建设提供借鉴。

3 国外人口健康领域科学数据共享平台建设的特点

3.1 拥有较完善的法律法规和政策体系

资源共享的概念早在公元前1世纪就被西方学者提出。近现代,发达国家的科技资源共享从

图书、科技文献的馆际互借开始,后扩展到中大型科学仪器设备共享,进而到自然资源科学数据和科学数据的共享。发达国家通过制定法律来保障科技信息的公开和共享,如美国的《信息自由法》、《美国联邦信息资源管理法》,德国的《信息和通讯服务规范法》,俄罗斯的《联邦信息、信息化和信息保护法》,法国的《信息社会法》^[2]等。1990年,美国颁布了《全球变化研究法案》,进一步规范了实施科学数据共享的研究项目类别,完善了相关数据整合的规范。90年代后期,美国政府建立了以“完全与开放”的共享国策为核心的法律和制度保障体系,标志着美国科学数据共享进入较为成熟的阶段^[3]。通过法律形式建立起科学数据的共享制度,并逐步形成规范、系统的制度体系,使得围绕科学数据收集、存储、管理、利用而建立的共享平台拥有了制度保障,有效地促进了发达国家科学数据共享平台的发展建设。

3.2 集成度高,利于数据深度挖掘和知识获取

随着科技发展和理念提升,发达国家建设的众多科学数据共享平台集成度不断提高,从最初实现数据的收集、存储,逐步完善、提升为集收集、存储、管理、整合、处理、分析、展现、应用于一体的综合性数据共享平台。

NCBI是美国国立生物技术信息中心建立的同名数据共享平台,建立之初主要是负责保管GenBank的基因测序数据和Medline的生物医学研究论文索引数据。经过多年发展,NCBI不仅建立了高效管理基因组、蛋白组、化合物等多种类数据的数据库管理系统,而且建立了与科技文献资源(如Medline)、临床数据资源(如Clinical Trials.gov)、公共卫生数据资源(如Pathogen Detection Project)等的紧密关联。NCBI还开发集成了在线数据综合检索系统Entrez,其在多个数据库间建立起可靠、完善的关联,可根据要求检索不同类型的相关数据和信息,如从查询一个DNA序列开始,检索到对应蛋白产物及3D结构图,再到相关文献,每个条目还给出与查询条目接近的信息,检索结果可以多种格式输出,也可

打包或逐个下载。为便于数据深度挖掘和分析,NCBI针对不同数据特点和应用场景开发了一系列工具软件,如基因序列比对分析工具BLAST、蛋白序列进化树分析绘图工具CDTree、用于DNA位点识别的电子克隆工具e-PCR、自动检测基金剪切位点工具Splign等,系列工具软件的开发及应用使研究人员通过平台得到了从数据获取到数据解读的一站式解决方案。当前,NCBI已综合集成了数据的上传、下载、检索、在线处理、整合分析等功能,同时还承担了新闻发布、在线培训、科研合作等职能。高度集成的共享平台使从数据到知识的过程进一步缩短,数据的科学价值、社会价值和经济价值更容易被挖掘。

3.3 开放共享程度高,数据来源广、更新快,用户群体宽泛

平台的开放共享程度体现了对数据的利用水平,也影响着平台自身发展的核心竞争力。国外人口健康领域科学数据共享平台十分注重平台资源的开放共享。为实现在世界范围内合作收集生物技术信息,共同开展生物信息研究,NCBI采取了科学数据完全开放共享的发展策略。如其最主要的数据库GenBank,任何个人或组织都可以在不需要注册的情况下通过平台查询并下载自己感兴趣的核酸序列数据。完全开放共享的发展理念和策略产生了数据资源的“虹吸效应”,使得NCBI的数据来源和用户群体遍布世界。NCBI的GenBank还与欧洲EMBL-EBI数据库、日本DDBJ数据库组成国际核酸序列数据库合作联盟,几乎收录了世界上所报道的所有核酸序列数据,并且每天实时更新交换各自的序列信息^[4]。

同样的情况在CT和UptoDate中也有体现。CT是目前国际上最重要的临床试验数据共享平台之一,由美国国立卫生研究院和美国食品药品监督管理局共同开发,主要向医学科研机构和人员提供临床试验的注册服务^[5]。通过注册,有效增加临床试验信息的透明度,增强了试验过程的规范性和结果的可信度。CT面向全球开放,截止到2017年8月5日,平台已拥有在全美50个州及全球其他200个国家开展的251021项临床试

验研究及其结果的数据,而且这一数字随试验注册情况在随时更新变化,任何机构和个人都可以在CT免费注册和查询这些数据。UptoDate是荷兰威科集团开发的一个全球领先的基于循证医学原则的临床知识共享平台,为临床医师、药师提供即时、循证的临床医药信息。UptoDate由上万个临床诊疗专题构成,每个专题由领域内经验丰富的医师综合已发表的一系列权威研究证据进行撰写和编辑,给出疾病诊疗的医学知识和能够运用于临床实践的分级推荐意见。目前,UptoDate每日更新,其数据已覆盖24个专科,有1.05万多个专题,9700多条分级推荐意见,3万多张图表,160多个医学计算器,5600多篇药物专论以及42.5万多条Medline参考文献,被广泛应用于180多个国家的3.2万多家医疗机构^[6]。

平台的高度开放共享,带来的是数据来源广度的增加和更新速度的提高,并进一步助推平台的建设发展和数据的深度发掘利用。

3.4 注重数据质量,建立了规范化的数据处理流程和较完善的质控体系

数据质量决定了数据价值,国外人口健康领域科学数据共享平台十分注重平台数据的质量控制,根据自身拥有数据资源的特点建立了相对规范的数据处理流程和较完善的质控体系。

在CT平台,用户提交一份合格的临床试验注册方案,需要按要求填写已被格式化为12部分的几乎涵盖了临床试验各方面内容的电子表单^[7],具体包括:(1)研究方案名称和背景资料,如各类标识号(ID、次级ID)、研究名称(精简名、缩写名、官方名)、研究类型(干预性研究、观察性研究、拓展性应用);(2)美国FDA相关信息,如IND(Investigational New Drug Application,临床试用新药申请)、IDE(Investigational Device Exemption,临床器械研究豁免)序列号;(3)受试者评审信息(评审委员会、数据督查委员会和监督当局信息);(4)组织者信息(试验责任方、主办方、合作方信息);(5)研究方案说明;(6)试验状况说明(核查日期、招募状况、拓展性应用状况);(7)研究方案设计;(8)

分组和干预; (9) 研究对象和关键词; (10) 受试者选择 (目标人群来源、抽样方法、纳入排除标准、性别、年龄限制、是否接受健康志愿者等); (11) 研究方案分中心及研究者信息; (12) 其他相关信息 (参考文献、相关网络链接等)。提交的信息将由其内在质控系统进行审核, 以确保数据准确、真实。UptoDate提供的是基于循证的临床医药信息, 为了给医务工作者的临床实践提供科学、有效的指导和参考, 其相关信息要经过7步严格的处理流程^[8]: (1) 收集循证证据和用户反馈意见; (2) 对证据开展评价与分级; (3) 进一步分析分级后的证据; (4) 如有循证证据被推荐, 对推荐建议进行分级; (5) 增加证据分级及推荐分级信息; (6) 专家审核; (7) 信息更新。在NCBI平台, 因生物信息数据体量往往达到GB、TB量级, 为避免数据的重复或不准确, NCBI还专门推出RefSeq数据库, 每个位点挑出一个代表序列来减少重复, 经过校正和冗余数据筛选后, 进一步提高了数据的可信度。

国外知名数据共享平台在数据质量控制方面还有不少方法值得借鉴。如将同行评审作为数据共享利用的前提步骤; 再如科学数据的分层管理^[9] (研究型数据、资源型数据、参考型数据)。通过制定一系列引用规范、标准和原则, 实现对科学数据全生命周期的质量控制。这些数据平台使用或推出的标准往往还是国家或行业标准, 如NCBI的GenBank、欧洲EMBL-EBI、日本的DDBJ为便于每日数据的交互, 有着较为统一的数据标准, 这样极大地增强了数据的可信度, 也进一步激发了平台活力。

3.5 服务类型多样, 提供个性化的数据服务

服务能力是数据共享平台的重要评价指标, 国外人口健康领域科学数据共享平台不断丰富服务模式, 以满足不同用户的个性化需求。

NCBI平台不断丰富功能使其服务能力得到持续提升。除了一般的浏览、检索服务以及为用户存储、管理生物信息数据的基础服务以外, 平台提供的一系列工具软件可以为用户在处理诸如序列数据检索、序列同源性和相似性对比、序

列多重排比、进化树分析、蛋白功能分析、大分子结构分析与多维显示等复杂生物信息数据分析时, 提供轻松、便捷的解决方案。在个性化服务方面, NCBI推出了“My Ncbi”。用户通过设置, 可定制符合自我偏好的检索过滤器, 实现快速准确检索, 并根据需要保存检索历史; 利用RSS (Really Simple Syndication, 简易信息聚合) 技术, NCBI还会根据用户特性, 向其实时推送研究相关的、最新的科学数据、文献、临床试验、临床治疗指南及其他相关数据库中数据的更新情况, 使用户能够及时掌握自身感兴趣的相关研究的最新动态。此外, 生物信息方法学相关的培训课程也是其重要的服务内容之一。UptoDate也十分注重用户的个性化需求, 其给出的分级意见可以满足不同层次医务工作者的知识获取需求, 还提供继续教育的服务; 平台内置的临床医学计算器可以帮助医务工作者在实施临床诊疗、疾病风险评估过程中方便、快捷地计算出关键指标数值。同时, UptoDate的医学主题内容还有患者专用版本, 可以为患者或普通大众提供个性化的、免费的医疗咨询和健康教育服务。

多样化和个性化的数据服务增强了平台的粘度, 服务的过程使数据价值得以体现, 数据的价值也在服务过程中得到了进一步提升。

4 国内人口健康领域科学数据共享平台建设现状与问题

国内的平台建设起步相对较晚, 近些年, 随着人口健康领域信息化基础设施建设水平的提高, 各类数据平台相继建立且发展迅速。当前, 在国内人口健康领域较具代表性、建设相对成熟、共享程度较高的数据平台是国家人口与健康科学数据共享平台 (NCMI), 也是国家科技基础条件平台下科学数据共享平台的重要组成部分。NCMI的前身是医药卫生科学数据共享网, 于2003年作为科技部科学数据共享工程重大项目立项, 主要完成对国家科技计划项目中人口健康领域科学数据的汇交、处理、存储、管理和共享等任务。2010年, 项目通过科技部和财政部组织的

平台认定转为长期运行。2016年，首次向社会公众发布了平台拥有的49.1TB、2.8亿条人口与健康大数据资源，涉及生物医学、基础医学、临床医学、公共卫生、中医药学、药学、人口与生殖健康七大类237个数据集^[10]。经过10余年的建设发展，NCMI的资源种类不断丰富、资源数量不断增多，并且在卫生决策、重大工程、科技项目、突发应急事件和服务民生方面取得了显著成效，但仍有不少需要改进、完善的地方，也较明显地反映出国内人口健康领域科学数据共享平台的建设现状及存在的一些问题。

4.1 法律法规缺失，制度保障体系尚待完善

目前，国内尚无科学数据共享方面的法律，仅在公共信息领域有一些行政法规、政策文件和部门规章，如《中华人民共和国科学数据共享条例（建议稿）》、《国家科技计划项目科学数据汇交暂行办法》。与国外科学数据共享规范、成熟的法律法规体系相比，我国相关法律法规的制定整体相对滞后^[11]。现有的一些行政法规和部门章程，在实际使用中原则性规定较多，缺乏可操作性。在对参与数据共享的机构和个人应用何种机制、如何向数据共享者返还公平的惠益方面，仍有明显缺失；更多的是从注重数据保护角度出发，对于共享利用方面仍旧欠缺。特别是随着人口健康领域的科学数据与公众个体的关系愈加紧密，制度保障体系的不完善，对数据共享平台的长期稳定建设和高效运行有较大影响。

4.2 集成度有限，功能亟待拓展

受制度保障体系、建设理念先进性、技术开发水平或其他方面的影响，国内人口健康领域科学数据共享平台的集成度十分有限。很多数据平台的建设仍停留在对数据库的基础操作水平上，如对平台数据的查询、下载，极少有平台能够像NCBI那样，将高相关度的科学数据、文献、临床试验、临床治疗指南等信息进行深度整合并及时推送，“支离破碎”的数据严重影响了数据挖掘和知识形成的效率。另外，人口健康领域的科学数据种类繁多、形式各异，对于一些特殊格式的数据，很多平台选择为其提供专用工具，却使

数据的读取和后续处理分析工作变得复杂，降低了用户的使用体验。国内数据平台亟需加强数据的整合集成，进一步拓展平台功能。

4.3 数据来源相对单一，更新较慢，质量参差不齐

国内的数据共享平台还普遍存在数据来源单一，更新较慢的问题。如“十二五”之前，NCMI的数据来源主要是“973”“863”、科技支撑计划等国家科技计划课题中产生的科学数据。虽然数据种类涉及到人口健康领域的各个方面，但因其他来源途径有限，数据的深度和广度较易受国家科技计划调整的影响，数据的补充、更新在课题结束后也变得十分困难。同时，因无相对统一的元数据标准，数据的质量也参差不齐。如NCMI虽然制定并发布了元数据方案，但总平台和各分中心的元数据描述并不一致^[9]。此外，平台发布的数据中很多缺失了科学数据的获取途径、方法标准、仪器标准、实验对象纳入排除标准等关键信息，使得数据的可用性和科学性大打折扣。

4.4 开放共享程度有限，数据服务有待提高

在数据共享利用方面，国内数据平台差距巨大。虽然对数据共享必要性和意义的理解越发深刻，也倡导数据的开放获取，但在实际操作中，从平台获取数据的过程十分繁琐。如在NCMI中，展示的多为元数据，数据实体很难通过在线检索、下载的方式被获取，也存在个别数据集链接无法访问的情况，数据的可见性、可得性较差，降低了用户体验。在数据服务方面，国内平台更加注重数据存储和数据保护，数据服务内容和能力有限，个性化服务方面更显不足。

5 启示

5.1 健全法规制度体系，扩大数据来源，提高数据共享利用

通过法律制度保障科学数据的共享利用是发达国家科学数据共享平台建设成功经验之一。为保证我国人口健康领域科学数据共享平台的长期健康、稳定发展，必须加快科学数据共享法律法规体系的建立与健全，使科学数据共享有法可依，不断完善科学数据共享的管理机制，明确数

据共享参与主体的责权利,规范科学数据的管理、存储、开发、共享与利用。进而建立国家人口健康领域科学数据共享的网络体系,扩大数据来源,拓展用户群体,加强平台间的合作交流,提高数据共享利用程度,使人口健康领域的科学数据形成从采集、处理到共享利用的良性循环。

5.2 完善数据标准体系,加强质量控制,推动平台规范建设

数据标准的完善程度体现了数据共享利用的水平。针对人口健康领域科学数据种类多、类型复杂的特点,应在参考国际主流和自身现有标准的基础上,从国家或行业层面进一步细化标准种类和适用范围,建立既能高度协调统一,又能顾及人口健康各专业领域科学数据特点的数据标准体系。平台应加强对数据的全生命周期质量控制,聘请或建立独立团队,广泛开展对数据平台建设,特别是数据质量的评价监督,推动国内人口健康领域科学数据共享平台的规范化、标准化和国际化建设,扩大和提高我国人口健康领域科学数据的影响力与科研价值。

5.3 加强数据整合集成,创新服务模式,拓展平台服务能力

人口健康领域科学数据共享平台应加强数据的分类、整序,降低分散数据之间的链接壁垒,对关联数据进行表征和组织,如将分子层面、细胞层面、组织器官层面、个体层面的信息加以整合集成,再如将传统医学和现代医学的观点加以重构,从而构建更加科学、高效的知识发现和获取途径。在此基础上,不断丰富服务模式,如针对临床医务人员数据管理技能不高的情况,开发简便、快捷的数据处理工具,或开展科学数据管理技能培训;为规范和提高数据利用,开展数据规范使用培训,或制定发布引用标准;针对科学数据实验中可能涉及的医学伦理问题进行科学讲座;提供数据提交与保存服务、相关信息推送的个性化服务等。通过提供良好服务,与用户建立更加紧密互信的关系,不断扩大平台的知名度和影响力。

6 结语

实施人口健康科学数据共享,整合离散的海量医学科学数据资源,是信息时代科技发展的必然选择,是增强医学科技竞争能力的有效途径。当前我国人口健康领域科学数据共享平台建设已取得了一定的成就,但也有很多的不足和较大的提升空间,不断完善科学数据资源共享利用的体制机制,把握好数据共享在个人、机构、国家及社会公众利益之间的平衡点,不断丰富数据资源,持续拓展平台功能,坚持标准规范建设,以服务为导向,建设更加安全、可靠的人口健康科学数据资源,使其更好地为人民群众的身心健康和经济社会的全面发展提供有力保障。

参考文献

- [1] 司莉,邢文明.国外科学数据管理与共享政策调查及对我国的启示[J].情报资料工作,2013(1): 61-66.
- [2] 完颜邓邓,高峰.英美澳科学数据存储与共享平台建设现状调查及启示[J].图书馆建设,2016(3): 29-34.
- [3] 吴松强,沈馨怡,刘晓宇,等.发达国家科技资源共享的经验与借鉴[J].实验室研究与探索,2014,33(6): 139-143.
- [4] 田耕,刘炯辉,兰翎.NCBI网站及Genbank数据库介绍[J].国外医学分子生物学分册,2000,22(5): 317-320.
- [5] CLINICAL TRIALS.What is ClinicalTrials.gov[EB/OL].(2017-02)[2017-08-05].<https://clinicaltrials.gov/ct2/about-site/background#WhatInformationCanIFind>.
- [6] UPTODATE.全球No.1的临床诊疗知识库[EB/OL]. [2017-08-05].<http://www.uptodatechina.com/>.
- [7] 王洋,雷艳.美国临床试验数据库注册流程与填写要求[J].中国中医药信息杂志,2012,19(7): 2-4.
- [8] 司富强,丁武国,韦当,等.四钟循证医学数据库比较分析[J].中国循证医学杂志,2013,13(5): 612-615.
- [9] 汪俊.美国科学数据共享的经验借鉴及其对我国科学基金启示:以NSF和NIH为例[J].中国科学基金,2016(1): 69-75.
- [10] 孟祥艳.国家人口与健康科学数据共享平台数据资源在京发布[J].中国科技资源导刊,2017(1): 封底.
- [11] 刘润达,彭洁.我国科学数据共享政策法规建设现状与展望[J].科技管理研究,2010(13): 41-43.