DOI: 10.13376/j.cbls/20240159
文章编号：1004-0374(2024)11-1311-04

睿 ● 观 ● 察

大科学设施的科学数据汇交与开放共享面临的挑战与对策

吴家睿[1,2,3]

(1 中国科学院上海高等研究院国家蛋白质科学研究(上海)设施，上海 201210；2 中国科学院分子细胞科学卓越创新中心，上海 200031；3 上海交通大学安泰经济与管理学院/主动健康战略与发展研究院，上海 200030)

国家重大科技基础设施 (简称大科学设施) 是国之科技重器；它们聚焦科学技术前沿，具有科研、技术、工程三重属性，体现了国家整体研究实力和创新能力。自 20 世纪 70 年代我国开始建设兰州重离子加速器和 80 年代建设北京正负电子对撞机，我国在大科学设施的建设方面有了长足的进步；截至 2023 年底，我国布局建设 77 个大科学设施，其中在建和运行的大科学设施达 57 个，已经进入了全球拥有大科学设施的“第一方阵”国家梯队。

按照中国科学院的分类标准，大科学设施分为三类：第一类是为基础研究、应用基础研究和应用研究服务的公用实验设施，如“上海同步辐射光源”(简称上海光源) 和“国家蛋白质科学研究 (上海) 设施”(简称蛋白质设施)；第二类是为国家经济建设和社会发展提供技术支撑的公益科技设施，如“中国地壳运动观测网络”；第三类是为特定学科领域的重大科研目标服务的专用研究设施，如“500 米口径球面射电望远镜”(简称中国天眼)。

作为以国家投入为主建设和运行的大科学设施，其关键特征是“开放共享”，如美国能源部向用户开放的 28 个大科学设施，每年大约为 70 多个国家或地区的科研人员提供技术服务。这一特征在前两类大科学设施上尤为突出：例如，上海光源自 2009 年 4 月投入运行至今，已经有 46 个实验站对国内外科研用户开放，每年向用户提供 4 000 至 5 000 个机时，仅最近 5 年就为 7 千多项研究课题提供了 2 万多机时服务；又如蛋白质设施，自 2015 年 8 月投入使用至今，已经为 430 家国内外用户单位的 1 万多项研究课题提供了实验服务，仅 2023 年就服务了 227 家单位的 1 396 项研究课题。

显然，大科学设施在为众多用户提供服务的过程中产生了海量的科学数据，如上海光源一年的运行服务就可以产生 20 PB (1 PB = 10^{15} Byte) 左右的科学数据。一般来说，大科学设施产生的科学数据主要有三类：从设施的科学装置或仪器设备上直接获取的原始数据；通过用户分析或加工原始数据而产生的研究数据；以及帮助研究人员理解数据背景和实验设计等与具体实验相关的元数据 —— 如实验用户、实验条件和环境等信息。在大数据时代，尤其在当前人工智能 (AI) 技术蓬勃发展时期，科学数据成为了赋能科学创新和技术创新的关键要素。政府资金资助的研究工作所产生的科学数据有两个重要的来源：一是大科学设施，二是科技重大项目。国家数据局等多个部门在 2023 年联合颁布了《“数据要素 ×”三年行动计划 (2024—2026 年)》，在第九条“数据要素 × 科技创新”里明确要求：“推动科学数据有序开放共享，促进重大科技基础设施、科技重大项目等产生的各类科学数据互联互通”。当前国际科技界对大科学设施产生的科学数据之开放共享也很重视，如位于瑞士日内瓦的欧洲核子 研究中心 (简称“CERN”) 专门打造了将其实验数据保存并供广大研究人员开放共享的“开放数据网站”(Open Data Portal)；该数据平台不仅开放给合作伙伴和用户，而且也开放给非合作的科研人员乃至普通公众。

虽然国内的大科学设施在总体数量和技术水平上都达到了国际水平，部分大科学设施的技术水平甚至为国际领先，但总体来说，在对大科学设施产生的科学数据之管理方面还很不完善，尤其在科学数据汇交和开放共享两个关键环节仍然面临着巨大的挑战。本文将围绕这两个方面的挑战进行分析，并提出相应的对策。

收稿日期：2024-10-08
基金项目：中国科学院战略性先导专项“多维大数据驱动的中国人群精准健康研究”(XDB38020000)
通信作者：E-mail: wujr@sibs.ac.cn

1 大科学设施的科学数据汇交面临的挑战与对策

将众多科研单位和研究人员在科研工作中产生的科学数据进行汇交是形成数据要素的关键举措。但在上个世纪“小数据时代”，各类科学数据通常都分散在不同的个体研究者领导的实验室，形成“数据孤岛”；这个特征在生命科学研究领域的“PI”模式下尤为突出。如何实现研究领域的科学数据汇交成了当今大数据时代需要解决的一个关键问题。在国务院2018年印发的《科学数据管理办法》里就明确规定了科学数据的“强制性汇交”原则：“各级科技计划(专项、基金等)管理部门应建立先汇交科学数据、再验收科技计划(专项、基金等)项目的机制；项目/课题验收后产生的科学数据也应进行汇交”。值得强调的是，科学数据的“强制性汇交”并非易事。因此，在该管理办法颁布7年后的今天，政府有关部门提出了更明确和更严格的规定，如国家发展改革委在2024年7月发布的重点专项“主动健康和人口老龄化科技应对”的项目申报指南中明确要求：“申报单位和个人应签署具有法律约束力的协议，承诺项目产生的所有科学数据无条件、按期递交到国家发展改革委指定的平台，纳入国家生物数据中心‘1+N’体系，在本专项约定的条件下对专项各承担单位，乃至今后面向所有的科技工作者和公众开放共享。如不签署数据递交协议，则不具备承担本专项项目的资格；签署数据递交协议后但未在商定的期限内履行数据提交责任的，由专业机构责令整改，拒绝整改者，专业机构追回项目资金，并予以通报”。

然而，我国政府和专业机构在大科学设施的科学数据汇交工作方面却没有做出相应的规定，如中国科学院在2019年发布的《中国科学院科学数据管理与开放共享办法(试行)》里，对大科学设施为用户服务所产生的科学数据之收集和保存明确提出需要与用户进行协商：“对于用户使用上述设施产生的科学数据，可在确保用户权益的基础上，通过协议的方式开展科学数据的收集和保存等工作”。可以说，虽然科研单位的用户通过政府公共资金资助在大科学设施服务产生的科学数据从理论上具有强制性汇交的属性，但现实中却并没有制度支持对大科学设施为科研用户服务产生的科学数据进行强制性汇交。因此，我国多数大科学设施为科研用户开展研究服务所产生的科学数据基本上没有进行强制性汇交，通常是由用户自身对其科学数据进行保存和使用，从而产生了众多的“数据孤岛”。这一问题在公用实验设施类的大科学设施上更为突出。也就是说，尽管公用实验设施类的大科学设施通过为众多科研用户提供实验服务而产生了海量科学数据，但是这些科学数据都散落在各个用户自己手中，大科学设施并没有统一保存这些科学数据，更谈不上开放共享这些科学数据。

由以上的分析可以看出，国家有关管理部门和专业机构应该制定相应的管理办法，明确把大科学设施为科研用户服务产生的科学数据纳入强制性汇交的范畴，以便能够更全面地汇聚来自科研项目和大科学设施的各类科学数据，从而在国家层面形成完整的科学数据要素。需要强调的是，大科学设施和科技重大项目是两个科学数据汇聚最主要的平台，前者基于大科学设施的装置或仪器设备，后者基于项目的总体部署和相关课题之间的内在联系。国家最近颁布的《“数据要素×”三年行动计划》正是这样进行设计的：“促进重大科技基础设施、科技重大项目等产生的各类科学数据互联互通”。

在为大科学设施的科学数据制定汇交管理办法时，重点要处理好“公共型科学数据”和“市场型科学数据”之间的关系。前者基于用户获得的各种政府资金资助的研究项目，后者则源于用户的企业资金或其他类型社会资金资助的研究项目。在国家2022年颁布的《关于构建数据基础制度更好发挥数据要素作用的意见》(简称“数据二十条”)的第三条“探索数据产权结构性分置制度”里明确指出，要“建立公共数据、企业数据、个人数据的分类分级确权授权制度。根据数据来源和数据生成特征，分别界定数据生产、流通、使用过程中各参与方享有的合法权利”。显然，大科学设施产生的“公共型科学数据”大多属于强制性汇交的范畴，而“市场型科学数据”则应该在保护企业用户或个人用户权益的基础上通过协议的方式进行非强制性汇交。要针对后者的市场属性制定相应的汇交办法，从而在大科学设施上实现《科学数据管理办法》提倡的目标：“鼓励社会资金资助形成的其他科学数据向相关科学数据中心汇交”。

2 大科学设施的科学数据开放共享面临的挑战与对策

科学数据最主要的价值体现在开放共享。可以说，科学数据的开放共享程度越高，其价值就越大。

换句话说，如果没有开放共享这一环节，科学数据汇交的意义就体现不出来。因此，《科学数据管理办法》里不仅规定在政府资金资助下产生的科学数据要实行强制性汇交，而且还明确提出：“政府预算资金资助形成的科学数据应当按照开放为常态、不开放为例外的原则”。这一原则在2023年国家颁布的《“数据要素×”三年行动计划》里也得到了重视，其中第九条“数据要素×科技创新”的第一句话就是：“推动科学数据有序开放共享”。

然而，包括科学数据在内的公共数据开放共享现状并不是很理想。国家数据局刘烈宏局长对此有过评述：“总体上看，我国公共数据的开放程度和利用水平，与社会各界期待相比，仍有很大差距。大家普遍反映，数据共享开放阻力大、顾虑多，数据供给的规模和质量都不够，资源利用的渠道和方式不丰富、不便捷”。让公共数据实现开放共享的主要制约因素是数据确权。为此，国家数据局前不久提出了推动公共数据资源开发利用的六项工作，其第一项工作就是要落实产权分置制度，明确公共数据授权运营的合规政策和管理要求，厘清数据供给、使用、管理的权责义务。这也正是“数据二十条”提出的主要任务之一，即“建立公共数据、企业数据、个人数据的分类分级确权授权制度。根据数据来源和数据生成特征，分别界定数据生产、流通、使用过程中各参与方享有的合法权利”。

一般说来，科学数据主要涉及到二种类型的权利：数据的所有权和数据的使用权。对于使用国家预算资金资助产生的科学数据而言，“强制性汇交”就意味着这类科学数据为国家所有，因而数据的开放共享就由国家来决定。这二者的关系在最近国家发展改革委的重点专项申报指南中就能清楚地看到：“申报单位和个人应签署具有法律约束力的协议，承诺项目产生的所有科学数据无条件、按期递交到国家发改委指定的平台，纳入国家生物数据中心‘1+N’体系，在本专项约定的条件下对专项各承担单位，乃至今后面向所有的科技工作者和公众开放共享”。显然，如果能够把大科学设施为科研用户服务产生的科学数据纳入强制性汇交的范畴，那么大科学设施也就被赋予了科学数据的所有权，从而可以决定科学数据的开放共享。

这里需要关注一个细节，用户对其实验产生的科学数据也有一定的所有权，需要得到保护。为此，国外许多大科学设施，尤其是光源类大科学装置，通常都有数据保护期的做法，即在实验结束产生了科学数据之后的一段时间里，数据只对该实验用户的成员或项目参与方开放；如位于瑞士的欧洲核子研究中心(CERN)下属的大型强子对撞机(LHC)的数据保护期为3年，3年过后数据才会公开让非用户的科研人员进行访问。也就是说，涉及到科学数据开放共享的基本原则是：制定开放共享的管理办法时要兼顾到用户知识产权之保护——正如中国科学院在2019年7月发布的《中国科学院重大科技基础设施运行管理细则》中所说：“在确保国家安全和保护知识产权的前提下，最大限度地实现科学数据共享”。如果仔细地推敲，这种“先用户，后公众”的科学数据保护与开放共享兼顾的策略在上文提到的国家发展改革委重点专项指南中也能看到类似的想法：“在本专项约定的条件下对专项各承担单位，乃至今后面向所有的科技工作者和公众开放共享”。

由此可见，科学数据的所有权和使用权紧密相关，应该先处理好数据的所有权，再有针对性地制定相应的开放共享规定。然而，大科学设施的科学数据有多种类型，其数据确权问题比较复杂。例如，从设施的科学装置或仪器设备上直接获取的原始数据应该是设施方和用户共享所有权；而对原始数据分析和加工得到的处理数据则取决于处理的方式，如果是设施的技术人员或相关分析软件等参与用户分析形成的处理数据也应该是双方共享权利，但如果是用户独立分析原始数据而获得的处理数据则是由用户拥有；此外，实验用户、实验条件和环境等与具体实验相关的元数据通常是由设施方拥有。

需要强调的是，大科学设施的科学数据之确权必须考虑数据的两个基本属性：即基于政府资金资助的研究项目产生的“公共型科学数据”，以及基于社会资金资助的研究项目产生的“市场型科学数据”；前者属于国家强制性汇交，而后者则是用户与设施通过协议的方式进行的非强制性汇交。显然，后者的开放共享方式首先是要考虑对企业用户或个人用户权益的保护。例如，一家生物制药公司“Captor Therapeutics”利用德国电子同步加速中心(DESY)的科学装置“PETRA III”进行了一个靶蛋白与相关配体的复合物之晶体结构分析，其数据为双方的共有产权而不对外开放。

也就是说，针对大科学设施产生的不同科学数据类型，需要从数据访问对象和数据保护期等角度对数据的开放共享制定有针对性的管理办法，如原始数据或元数据一般情况下只对利益相关方或者注

册的用户开放。对于涉及个人隐私和生物安全的医疗健康数据之开放共享还需要考虑个人信息的保护。显然，大科学设施产生的科学数据之开放共享需要考虑到数据的所有权和使用权之统一性，以及数据安全和数据开放之间的平衡关系。这方面可以参照国家的相关规定——“数据二十条”之第四条“推进实施公共数据确权授权机制”明确提出：“对各级党政机关、企事业单位依法履职或提供公共服务过程中产生的公共数据，加强汇聚共享和开放开发，强化统筹授权使用和管理，推进互联互通，打破‘数据孤岛’。鼓励公共数据在保护个人隐私和确保公共安全的前提下，按照‘原始数据不出域、数据可用不可见’的要求，以模型、核验等产品和服务等形式向社会提供”。