关于“共享公共数据 实现科学数据自立自强”的提案

第一提案人：陈松蹊

联名委员：毕彦超、陈增敬、郭媛媛、焦念志、金李、刘晓梅、舒勇、苏洵、唐冬生、王怀民、王亮、张甘霖、张云泉（征集到联名委员共13人，按姓氏字母序排列）

提案正文：

2023年2月，中共中央、国务院印发了《数字中国建设整体布局规划》。《规划》明确提出要畅通数据资源大循环，构建国家数据管理体制机制，健全各级数据统筹管理机构。推动公共数据汇聚利用，建设公共卫生、科技、教育等重要领域国家数据资源库。以数字化驱动生产生活和治理方式变革，用数据生产力为中国式现代化建设注入强大动力。

公共数据是指国家授权的公共管理或服务组织，收集、产生的涉及公共品并不含个人隐私的数据。公共数据具有公共性和非竞争性特征，通常包括各类地理空间、气象、大气环境、生态、流行病、经济、农业、交通、人口和社会数据等。

目前，数据驱动的研究范式正在深刻改变科研生产力。公共数据作为重要的科技资源，是众多科技领域，如人工智能、大气环境、统计学、医疗健康和经济管理学等，在解决国家重大需求、卡脖子问题中所必需的研究基础。

目前我国科技工作者在获取公共数据上面临诸多困难，具体表现为以下三个方面。

1.公共数据获取渠道不畅。近十年来，我国一些公共数据的可获取性得到了提升，一些数据的实时播报为通过网络实时下载数据提供了可能。但是网络下载无法获取历史数据，下载数据的通道并不稳定，数据格式时有变化，易造成数据缺失，研究成果的数据源容易被挑战。目前一般科技工作者缺乏国内历史公共数据的有效获取途径，而提供公开下载的数据来源是科研发表的基本要求。

2.科学研究过度依赖国外公共数据集。由于国内公共数据获取困难，中国科学家大量使用国外的公开数据集进行科学研究。经常使用的数据集有英国生物银行基于大样本人群的遗传、生活环境和健康数据；欧洲中期天气预报中心发布的自 1951 年的高分辨率全球气象再分析数据；美、欧、日本等机构发布的涵盖大气化学要素、二氧化碳、沙尘、灯光等高分辨卫星数据；世界卫生组织发布的各国流行病数据等。过度依赖外部数据，不利于我国科学技术自立自强，可能会限制研究人员的自主性和创新性；不利于掌握科技资源的主动权，存在关键时刻数据获取中断的风险；也不利于我国科技工作者讲好中国故事。

3.缺乏高质量的再分析科学数据集。观测数据普遍存在空间分布不均、时间延续性差、观测种类不全等缺陷。“再分析数据”作为多源数据的标准化集合在科学研究中起着关键的作用。再分析数据使用先进的统计方法，将物理模型与多源观测数据进行融合，是现有技术条件下的最优数据集。再分析数据的构造高度依赖稳定的数据源与职能机构有序公开的数据政策。目前再分析数据是人工智能算法训练的数据基础，华为盘古气象大模型就是基于欧洲气象中心公开的再分析数据集训练成功的。

针对上述问题，我们提出两点建议：

1.**按照数据风险等级，有序开放共享公共数据。**有序开放共享公共数据能够使国内科研人员、企业及时获取长时期历史数据，提高我国大数据分析和数据赋能能力。我们建议，高分辨率气象、大气、环

保、生态、经济社会等不涉及国家安全的数据应优先考虑公开。对一些敏感数据，可以签署标准化协议，对数据的使用进行不同程度的规范，之后再对国内学者和企业开放。

2.**集中力量打造高质量再分析数据集**。建议组建由领域与数据科学家组成的数据融合团队，发挥我国在数据同化方面的统计学基础优势，在一些关键科学领域构建高质量的再分析数据集，解决我国科研人员的数据需求，降低对外部数据的依赖，实现科学数据自立自强。

综上，我们建议优化有关政策，实现公共数据的开放共享，也促进非公共机构的数据流动和价值创造，引导中国数据要素市场的建立和良性循环。高质量科学数据集将为中国科学数据的自立自强奠定基础，为数据赋能科学研究和经济发展提供数据支撑。