开放科学数据，助推科技创新

——中国科学院建成开放服务的科学数据云

□ 本报记者 李建荣 通讯员 张杨

迅速发展的信息技术正不断助推科研行为方式的变革和科技创新发展。当前，世界各科技强国已经把科研信息化作为21世纪科技创新的战略举措。在我国，科研活动信息化已是提高科研水平和创新能力的必要手段。

作为中国科技的“国家队”，中国科学院（以下简称中科院）一直高度重视科学数据在科研发现、信息化建设中的创新及应用。上世纪70年代，中科院开始建设专业数据库。1982年科学数据库被列入中科院“七五”和后十年的10项重大基本建设项目。1986年国家计委正式批复同意建设“中国科学院科学数据库及其信息系统”，1987年科学数据库数据资源和信息系统正式启动建设，1997年获“中国科学院科技进步一等奖”，1998年获“国家科技进步二等奖”，基本形成了以研究所和课题组自主自治为单元的科学数据资源建设和积累模式。“十五”期间，科学数据库建设逐步系统化、规范化，共建成503个专业子库。“十一五”期间，在中科院信息化专项和国家科技基础条件平台等支持下，科学数据库逐步形成结构合理的科学数据网格体系，整合可共享数据量达148TB。

“十二五”期间，随着传感器、信息获取等数字技术的不断发展，科学数据也以史无前例的速度急剧增长。面向科技创新和科研信息化新需求，中科院启动“科技数据资源整合与共享工程”建设。“科技数据资源整合与共享工程”涵盖数据存储与管理云服务环境、海量科学数据分析与应用示范、科学数据整合与共享服务等三个子项目，工程着眼于“海·云”服务思想，开展海量存储基础设施服务、海量数据资源共享服务和数据密集型公共支撑服务，全面推进数据环境建设和持续深化数据应用，成为立足中科院，面向科技界，共享开放、服务创新的国家级科技数据中心。

在中科院的统一部署推动、全院50多家下属单位共同参与下，中国科学院计算机网络信息中心作为科学数据库牵头建设和技术支撑单位，紧紧抓住信息技术发展的脉搏，推动科学数据库在建库、整合和应用的全方位成长。科学数据库践行由硬件建设向环境构建、工程化项目向持续化发展方针，以云服务模式为基础，形成支持科研活动与科技创新的数据云，并从基础设施、数据资源、应用平台三大类服务的角度整合集成各类资源和服务，形成中国科学院数据云环境。

从最早“七五”期间15家单位、21个数据库，发展到目前“十二五”期间58家单位、1340个数据库，中国科学院数据云整合了从资源学科领域到植物学科领域等多领域数据库资源，提供共享数据量已从2.68GB增加到655TB，年均在线访问超过千万人次。“十二五”期间，共发表论文751篇，申请软件著作权55项、专利30项。累计为131项科研项目提供了数据支持和服务，在支持科研项目、支撑学科发展和服务经济社会发展等方面均取得良好的效果。项目积累的存储、处理与应用等资源整合为数据云一站式服务的相关技术，为持续推动科学数据云发展打下了坚实基础。

一、面向科研创新前沿 构建科研服务新模式

中科院数据云以数据资产为核心，充分利用先进的云计算技术，整合数据全生命周期的重要设施与资源，是现代科研创新体系的重要组成，是大数据科研成果服务于社会应用的示范平台。

中科院数据云环境为科研活动提供以海量存储设施为基础的云存储、云归档、虚拟机和数据云等服务，为科学数据管理和共享提供运行支撑环境，为科研创新活动存储提供了有效保障。截止到2015年，中科院数据云存储环境运行服务总容量达52PB，云存储规模达8PB，共拥有物理服务器约300台，虚拟机5000+的计算服务能力。数据归档总容量达38PB，拥有归档能力大于20TB/天，在线磁盘阵列容量达到2PB，近线磁带库存储容量达到30PB的归档系统。建成布局中科院、直达各所的“一主一备+12分中心”的分布式、可扩展存储系统，提供满足国标5级的“同城双中心”、“两地三中心”的高等级灾备服务。

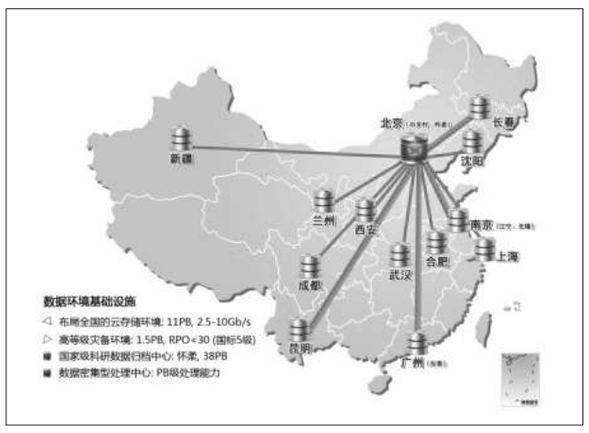

图1：中科院数据云“一主一备+12分中心”分布式、可扩展存储系统

“十二五”期间中科院数据云形成以基础设施云服务、科研数据云服务、数据应用云服务为主体的多层次、交叉式信息化服务体系。中科院计算机网络信息中心通过研发部署云计算系统，为中科院信息化专项、先导专项、重点基金项目、科技支撑计划等项目提供支持。以生物信息学分子数据分析环境、地理空间数据云、DViz大数据可视化等应用的开发，示范了多学科领域数据、模型及云服务应用的技术手段与服务模式。2015年8月，中科院网络中心成功申请并获批我国首批试点网络连续性出版物，创办《中国科学数据》期刊，探索建立科学数据产权保护的新方法，推动科学数据出版与数据引用，进一步促进我国科学数据资源的开放与共享。

在服务科研的同时，中科院数据云面向社会需求不断加强产业化创新服务，提升拓展技术优势。在交通管理、食品安全、新材料研发等公共领域，中科院计算机网络信息中心与国家发改委、食药监总局、北京地税等三十多家企事业单位开展相关合作。2012年获得中国产学研创新合作奖，2013年获批成立大数据应用服务技术北京工程实验室。2014年、2015年先后两年成功举办科学数据大会，吸引了来自全国科研院所、高校以及相关企业参加。

二、中科院数据云成果五大亮点

2015年8月31日，国务院发布了《促进大数据发展行动纲要》标志着我国正式把发展大数据上升为了国家战略。中科院数据云服务平台的建成，将进一步释放我国科学大数据价值，为“一带一路”、“生态文明”、“科学前沿”、“基础学科”与“创业、创新”等国家战略需求及社会热点应用提供了有力的数据支撑与科学技术应用服务。

（一）让中国科技照亮“一带一路”

实施并建设好“一带一路”，是融合中国发展优势与全球合作愿景，实现中国梦的一个重大举措和抓手，为促进区域共同繁荣和世界和平发展提供了新契机。“一带一路”建设需要科技创新引领和驱动，依靠科技创新支撑“一带一路”实现可持续发展已成为战略共识。

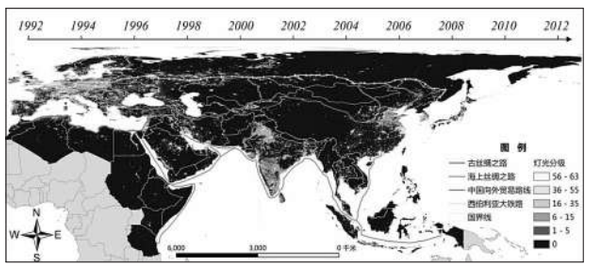

图2：1992—2012年夜间“一带一路”区域夜间灯光变化数据

1.大数据协同平台提供有力数据保障 成为主管部门决策“智库”

“一带一路”建设涉及新亚欧大陆桥、中蒙俄、中国—中亚—西亚、中国—中南半岛等多个经济走廊，经济带建设需求已对科学技术发出强劲召唤。2015年4月，中科院白春礼院长做出批示，支持并推动建设“一带一路”国际科学家联盟和信息网络平台。平台以中科院为主导，着力打造满足国际科学家了解问题、开放研讨、协同研究和信息共享需求的协同创新网络平台。

中科院信息化建设专项课题“资源学科领域基础科学数据整合与集成应用”以俄罗斯、蒙古等“一带一路”国家的基础地理与资源环境为本底资料，通过整合获取沿线国家的人口、经济、能源、交通设施等数据资料，集成大数据信息，直接为“一带一路”科学院联盟和协同创新网络平台提供数据，实现了为“一带一路”建设决策和国家治理提供长期的科技战略咨询作用。

2.环境监测数据服务于“一带一路”区域环境治理与资源开发

“21世纪海上丝绸之路”战略实施过程中，海上经济活动和海洋保障都需要海洋科技发挥基础支撑作用，而海洋数据作为海洋科技发展的基础，其有效管理及共享为国家战略实施提供重要的科学依据。中科院海洋研究所科研数据整合项目整合了包括观测浮标、航次调查、国内历史资料等多源数据，形成了集水上、水面、水下数据于一体的海洋立体综合数据集，特别是在中国黄海、东海，长期连续的观测数据与开放航次等调查数据组成的观测研究网络，为保障海上丝绸之路正常运行提供了基础海洋环境数据。此外，通过多源数据的整合，科研工作者也可更加方便地获取海上丝绸之路沿线区域的调查数据，推动海上丝绸之路沿线海洋资源的开发，创造更大的社会经济价值。

3.语言资源数据库推动“一带一路”区域文化与科技交流

中科院合肥物质科学研究院牵头负责的多民族语言资源数据库为“一带一路”少数民族地区的言语教学和言语科研提供了坚实的语言数据基础。数据库将藏语言语数据库应用于当地少数民族青少年的双语教学，促进当地的对外开放与合作。此外，将蒙语和维语的言语数据库嵌入面向少数民族地区的旅游信息产品中，将旅游领域的汉语日常会话翻译成少数民族的语言语音，加强游客对“一带一路”相关少数民族地区的了解，利于少数民族地区的旅游业发展。

中科院自动化研究所中文语言资源库建立了“100万词蒙语单语语料库”、“汉藏双语句子级对齐语料库”、“维吾尔语—汉语综合领域平行语料库”等语料库，增进少数民族语言地区与汉语普通话地区的学术交流，加强上述地区与“一带一路”上蒙语、藏语语言国家和地区的纽带作用，促进新疆、内蒙古等地发挥区位优势，提升其作为向西开放的枢纽和文化科教中心地位。

4.科学数据开放为国际科学数据引进和交流共享奠定基础

中科院地理科学与资源研究所“东北亚中亚地区资源环境科学数据共享培训班”在授课期间，以中科院资源学科领域的“人地系统数据库”作为数据共享教学资源，并由该平台资源建设、平台开发和标准研制人员授课。来自俄罗斯、吉尔吉斯、塔吉克斯坦、乌兹别克斯坦、哈萨克斯坦、蒙古、泰国、巴基斯坦、孟加拉国的29名青年科学家接受培训，在掌握资源学科领域科学数据共享的技术和方法的同时，也获得了国际相关区域科学数据资源，为进一步加强“一带一路”区域的国际科学数据引进和交换共享奠定基础。

（二）让科技创新成为美丽中国的绿色引擎

生态文明建设需要科技创新支撑和引领。当前以大数据为基础的新一轮科技革命和产业变革，对我国的绿色发展既是挑战，也是机遇。如何将科技创新作为战略基点，加快培育和发展新兴产业，推进传统产业优化升级，支撑引领绿色发展成为时下科技工作者的新使命。

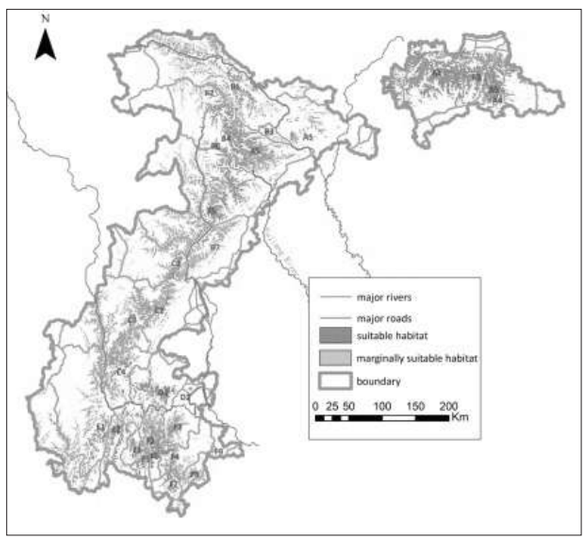

图3：大熊猫栖息地保护数据图

1.生态系统与安全数据库为全国生态功能区划提供依据

全国生态系统评估与生态安全数据库为全国和区域尺度的生态环境重大科研项目提供了数据支持，同时为国家生态环境保护、生态文明建设提供了重要科学支撑。由环境保护部与中科院联合颁布实施的《全国生态功能区划》以全国生态系统、生态服务功能及生态敏感性数据为基础。全国生态系统评估与生态安全数据库还为区域和地方生态保护与生态文明建设提供了数据支撑，在长江流域生态健康评估中，明确了长江生态环境状况、面临的生态环境问题与未来生态风险；在北京市生态保护红线规划研究中，明确了北京生态保护的关键区域；在内蒙古阿尔山市生态系统生态总值核算中，为地方开展生态效益核算开展了示范。

2.南海海洋科学数据库支撑我国海洋经济发展和海洋权益维护

党的十八大报告提出“大力推进生态文明建设”的战略决策并明确指出保护海洋生态环境。海洋是地球的主体，海洋生态子系统的状况对地球生态母系统起着举足轻重的影响，海洋生态文明是整个生态文明建设的重要方面。

立足南海，跨越深蓝。围绕热带海洋环境与资源两个重大研究方向，中科院南海海洋研究所南海海洋科学数据库致力于海洋动力环境与观测技术、边缘海地质演化与油气资源、海洋生态与生物资源优先学科领域科技数据资源的整合，南海海洋研究所数据资源体系和一站式共享服务系统的建设，支撑我国海洋科技创新、海洋经济发展和海洋权益维护。

3.地理与湖泊数据库为湖泊流域生态文明治理提供决策依据

湖泊流域大多为人口和经济发展密集区，流域生产生活排放的大量氮、磷等营养盐进入湖泊，造成湖泊富营养化和藻类水华频发，湖泊成为我国水环境问题最为突出的地理单元。目前，太湖、巢湖和滇池等大型湖泊富营养化突出，藻类水华暴发的水污染事故频繁发生。

围绕湖泊水环境保护，中科院南京地理与湖泊研究所承建的“南京地理与湖泊研究所数据整合与共享应用示范”开展了“面向政府决策的湖泊水环境治理决策与预警”专题服务，为太湖流域水资源保护局、巢湖流域管理局掌握太湖和巢湖蓝藻水华范围分布及水华面积，提供了及时有效的信息。在太湖、巢湖蓝藻调查、水资源调度以及流域水资源保护等方面起了较大的支撑作用，并为有关行政管理决策提供了依据，受到太湖流域水资源保护局的高度认可。

（三）取之于科学用之于科学 科学数据库激活科学前沿新研究

数据的爆发式增长，已把科学研究各个领域和环节推到了一个前所未有的“大数据”时代。一个国家的科学研究水平将越来越多地取决于其在数据的优势以及将数据转换为信息和知识的能力。中科院数据云作为科学大数据的基础数据库，在促进我国科学技术研究占领国际制高点上发挥了越来越多的支撑作用。

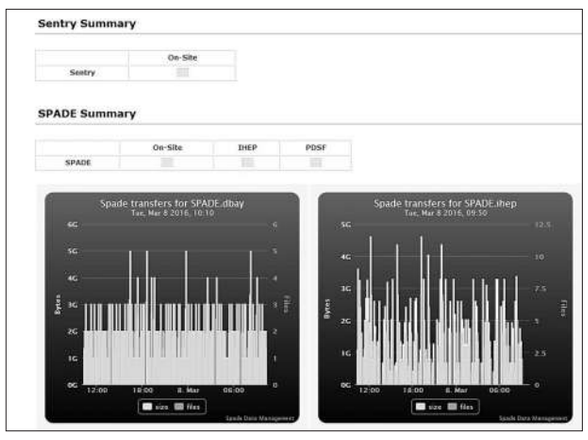

图4：大亚湾中微子实验数据传输

1.核能数据处理开启核能领域未来发展大门

大亚湾反应堆中微子实验是由中科院高能物理研究所主导、中美亚欧等国家和地区参加的大型国际合作项目，主要目标是利用核反应堆产生的电子反中微子来测定具有重大物理意义的参数—中微子混合角。中微子实验数据库主要存储大亚湾实验产生的实验数据，结合数据中心计算环境向大亚湾国际合作组的研究人员提供数据和计算服务。

中微子实验正式取数以来，取得了突破性的研究成果。2015年，大亚湾国际合作组在《物理评论快报》发表了中微子测量的最新结果，将中微子混合角θ13和中微子质量平方差的测量精度都提高了近一倍，为世界最高精度。大亚湾中微子实验获得的研究成果，开启了未来中微子发展的大门，产生了极大的社会影响。 2012年，首次精确测量θ13，入选美国Science杂志“2012年度十大科学突破”，为此大亚湾中微子实验合作组在2013获得“影响世界华人大奖”提名；2015年，大亚湾国际合作组发表迄今世界最精确的反应堆中微子振荡测量结果，实验团队获 “基础物理学突破奖”，这是中国科学家和以中国科学家为主的实验团队首次获得该奖项。

面向核能发展对数据的紧迫需求，为解决核能学科领域数据资源匮乏、分散等严峻问题，中科院核能安全技术研究所在中科院“十二五”信息化专项的支持下，联合中科院计算机网络信息中心、高能物理研究所、近代物理研究所等优势单位经过3年的持续建设，建成包括核数据库、核材料数据库、可靠性数据库、聚变数据库等数据资源，二十余套在线服务软件的综合性数据平台。核能数据库网站，已为来自中国、美国、英国等二十多个国家11500余名核能研究人员提供了核能数据及在线计算服务，用户累计下载量超过2TB，为核能设计及安全分析提供了全面的支持。核数据库子库HENDL面向先进核能系统核数据应用需求，成功解决了世界首个嬗变高放射性核废料ADS系统设计关键问题。核反应堆材料子库支持世界三大低活化马氏体钢之一的CLAM钢性能优化，为世界核材料领域低活化钢研发做出了突出贡献。

2.中国植物物种信息数据库开辟后植物分类学新时代

随着生物多样性信息学、新一代互联网技术的发展与应用，以及后基因组时代测序技术的发展，植物资源和植物多样性的研究遇到更多新的挑战。基于中国植物物种信息数据库基础上编著的《中国植物志》出版后，昆明植物研究所率先提出了“iFlora研究计划”。iFlora研究计划拟基于《中国植物志》的研究成果，整合植物学、分子生物学、生物信息学等现有优势学科力量，通过与生态学、自然地理学、植物化学、计算机科学等学科的交叉，打破传统意义上的纸本和单一产品的《植物志》的界限，实现植物物种多样性研究标准化、信息化和动态化，满足我国生物多样性保护研究与资源持续利用需求。“iFlora”研究计划的提出，开辟了后植物分类学的新时代。

（四）科学大数据孕育科研方法新范式

大数据作为改变人类生活及理解世界的新方式，正驱动着科学研究范式的转化，科学大数据已成为科学发现与知识创新的新引擎。从海量数据中解析所蕴含的新模式，科学大数据正带来科研方法论的新范式。

图5：近地天体望远镜

图6：太阳射电频谱仪

1.高能天体物理数据库成为我国空间天文科学体系中的重要组成部分

随着全球大型巡天观测项目的开展，天文学研究从小样本向着大数据模式转变，海量的天文数据给天文学家带来了巨大的机遇和挑战，天文学的研究也越来越离不开大数据集的统计分析，即数据挖掘和知识发现。

硬X射线调制望远镜（Hard X-ray Modulation Telescope，简称HXMT）卫星是我国正在研制的既可以实现宽波段、高灵敏度X射线成像巡天又能够研究黑洞、中子星等高能天体的短时标光变和宽波段能谱的空间X射线天文观测设备。HXMT将于2016年发射升空，并发布大量科学观测数据，用于开展致密天体和黑洞强引力场中动力学和高能辐射过程、X射线脉冲星的物理性质等方面的研究。

中科院先导专项项目“HXMT数据处理技术”将建成具备对HXMT卫星有效载荷实施在轨性能分析、完成数据处理与数据产品生成、提供数据发布与用户支持服务的数据分析平台，高能天体物理数据库为科学用户开展数据分析提供基础支撑，并成为我国空间天文科学体系中的重要组成部分。

2.海量土地数据确立我国土系变化趋势

在高强度利用下，我国农田究竟是丢碳还是固碳，国内外争论很多。在此之前，由于科研过程长时期缺失足够数据支撑造成结果难以定论。“中国农田土壤固碳潜力与速率研究”课题基于我国农田土壤有机碳采样分析和中国土壤数据库历史数据，进行“面对面”和“点对点”的比对，对于我国农田土壤碳库变化进行了研究。初步结果显示，除了东北地区丢碳，其他区域都有不同程度固碳。“中国土壤数据库”在该项目中提供了本底的土壤数据，对于土壤固碳速率正确估算，并确立我国农田主要是碳汇等结论提供了关键的数据支持。

面向《内蒙古自治区土系调查与<中国土系志·内蒙古卷>编制》项目的需求，中科院地理科学与资源研究所基于收集整理的原始数据、初级加工数据以及项目成果数据建立了内蒙东四盟土壤分析剖面实物和数据组。东北地理所黑土数据整合中心负责对课题采集的剖面数据和表层样点数据进行分析，并通过空间处理落实到相关图位上，建立土壤剖面实体模型，为中国土系的建立奠定了基础。

3.生物库成为科研人履行保护生物多样性公约的具体行动

生物多样性是人类共同的财富，也是人类社会赖以生存和可持续发展的基础。为了摸清中国生物多样性的家底，中科院生物多样性委员会自2007年起组织国内外100多位分类学专家，依据物种2000标准数据格式，每年编研、更新《中国生物物种名录》，并与全球生物物种名录实现信息共享。2015版《中国生物物种名录》，包括了动物界、细菌界、色素界、真菌界、植物界、原生动物界和病毒等七个部分，共收录物种8.3万个，编研过程中参考了中国动物志数据库，中国动物名录数据库、动物名称引证数据库。《中国生物物种名录》的编研和发布为生物多样性保护政策和规划的制定提供科学依据，为开展生物多样性科学研究提供基础数据，为公众参与生物多样性保护创造必要条件，是中国贯彻实施《中国生物多样性保护战略与行动计划》和积极履行《生物多样性公约》的具体行动。

（五）大数据撬动创业创新新应用

在信息经济发展迅猛的今天，大数据扮演生产要素的角色，让数据在碰撞中聚变，充分释放大数据的价值，带动“大众创业、万众创新”是中科院数据云的应用目标，虽然我国基于大数据的创业、创新业务和服务模式还不成熟，但却意味着更多机会，中科院数据云实际应用中已不断涌现出基于大数据的新尝试和探索。

1.灾种、救灾数据库为应急救灾提供灾害预测等创新服务

2014年10月，广东登革热疫情严重，为了支撑军事医学科学院的救灾防疫行动，“资源学科领域基础科学数据整合与集成应用”为其提供了广东省乡镇级数字地图、广东省面状人口数据和GDP数据、广东省土地利用数据直接应用于疫情聚集区的分析、重点采取防控区域的确定、传播风险的预测，为防疫救灾和危险评估提供了保障。

2015年4月，尼泊尔发生8.1级地震。“资源学科领域基础科学数据整合与集成应用”人地系统主题数据库迅速反应，第二天就整理出灾区及周边范围的基础地理、冰川冰湖、人口及社会经济、土地覆盖、历史地震资料等15个数据集，无限制、无偿向公众开放下载。通过开放尼泊尔数据直通车，快速集成不同灾种、救灾阶段所需要的数据资源和产品，为应急救灾提供无障碍的无偿共享服务，成为科学救灾的重要依据。

2.DNA条形码标准参考数据库助力森林公安快速破案

随着分子生物学的快速发展，DNA条形码为快速的物种鉴定提供了分子水平的精细分类学标准。该技术通过建立一套基于标准短基因片段的数字化序列文库来实现物种鉴定。

森林公安、海关等有关部门在打击野生动物盗猎、珍稀植物砍伐时，很多时候发现的是一些骨头，毛皮，甚至是一些木屑等不完整样本，而依法追责一定要鉴定出这些是动植物的具体信息。中科院昆明植物研究所在获得迪庆州森林公安的木屑标本后，通过与其建设的标准数据库进行比对，不仅鉴定出这些木屑来自红豆杉，而且准确地告诉了这些红豆杉大概生活区域，即采伐地。森林公安凭借这份鉴定报告，快速地破获了这起盗伐偷运案件。

3.语言资源库促进人工智能领域产品研发

中科院自动化所中文语言资源库项目在建立和整合语言资源的基础上，形成系列化的标准和规范，整合百余套数据库，建立了数据支撑服务平台，大大提高了语料库的有效获取和共享利用，并积极开展与企业合作，将语料库应用到企业的创新技术、新产品研发中。平台的数据库大量应用于30余个企业的技术研发，支持包括百度在内的商业公司的产品研发中。基于“语音合成语料库”等数据资源研发的语音合成技术，已与三星和联想分别合作，应用在其多款手机中。

“十三五”期间，在国家大数据行动背景下，以中科院“率先行动”计划为行动指南，面向智慧中科院发展愿景，中科院数据云将以科研需求为牵引，社会应用为落脚点，继续推动科学大数据的整合与开放、提高科学大数据为科学家与公众的服务，探索科学数据库发展和共享服务新模式。

科学大数据正在使科学世界发生变化，促进数据密集型科研范式的产生。中科院数据云先进的发展理念和有效的运行机制，有力的引导和整合了科学数据基础性工作，将科学数据战略机遇转化，成为数据密集型科学发现的制高点和前沿阵地。科技引领着社会的发展，面对“互联网+”、“万众创新、创业”的时代号召，科学大数据将释放出巨大潜力，在社会管理、民生保障、产业发展方面提供新的动能。