中国科学院科学数据云建设与服务

黎建辉，周园春，胡良霖，刘峰，朱艳华，沈志宏，吴章生，张杨
中国科学院计算机网络信息中心，北京 100190

摘要
科技数据资源整合与共享工程是中国科学院“十二五”五大信息化工程之一。总结了该项目的整体建设思想、建设情况、技术创新和服务创新等内容。截至项目结束，数据工程建成了存储容量达52 PB的分布式海量存储环境；整合可共享科学数据总量近655 TB，累计访问人次9 629万次，累计下载量456 TB；同时为用户提供强大的科学数据与文献互联以及丰富的可视化展示平台。工程实现了以基础设施云服务、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，逐渐建设形成共享开放、服务创新的国家级科技数据中心。

关键词
科学数据；数据平台；数据共享服务；服务成效

中图分类号：N37 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016061

Scientific data cloud construction and service of Chinese Academy of Sciences

LI Jianhui, ZHOU Yuanchun, HU Lianglin, LIU Feng, ZHU Yanhua, SHEN Zhihong, WU Zhangsheng, ZHANG Yang
Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China

Abstract
Scientific Data Resource Integration and Sharing Project is one of the 5 major informatization-specific projects of CAS for the 12th Five-Year Plan period. The overall construction of the project ideas, construction, technical innovation and service innovation, etc., was summarized. By the end of the project, a distributed mass storage environment with storage capacity of 52 PB was built. At the same time, it provided users with a strong connection between scientific data and literature and a rich visual display platform. The project has initially achieved a multi-level, cross information service system that included the infrastructure cloud service, research data cloud service and data application cloud service. It has gradually become a national science and technology data center for open sharing and service innovation.

Key words
scientific data, data platform, data sharing service, service effectiveness

1 引言

迅速发展的信息技术不断推动科研行为方式的变革和科技创新的发展。大数据在科研领域的蓬勃发展给科研方式带来了革命性的改变。作为大数据的重要组成部分，科学大数据驱动科学研究进入数据密集型科学发现范式这一全新阶段，已成为科学发现的新型战略资源[1]。为了抢占科技竞争的至高点，世界各国已纷纷把科学大数据纳入国家战略，并开始重点部署。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国家战略。

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。20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10年的10项重大基本建设项目。1986年中华人民共和国国家计划委员会正式批复同意建设“中国科学院科学数据库及其信息系统”，并于1987年正式启动建设。该项目1997年获得“中国科学院科技进步奖一等奖”，1998年获得“国家科技进步奖二等奖”，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的支持下，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148 TB。

“十二五”期间，中科院面向科技创新和科研信息化需求，启动“科技数据资源整合与共享工程”建设，目标着眼于“海·云”思想，全面推动全院科技数据基础资源、海量存储与处理基础设施、数据集成与应用先进环境的建设与服务。“科技数据资源整合与共享工程”项目涵盖数据存储与管理云服务环境、科学数据整合与共享服务、海量科学数据分析与应用示范3个子项目。截至项目结束，数据工程已建成了52 PB存储容量的数据资源中心，系统地整合了58家单位的科学数据库，可共享数据量达655 TB，重要数据服务130余例，在服务科技创新、国家战略、学科发展、社会应用、国际合作等方面发挥了重要应用。

2 “十二五”数据工程整体建设思想

中国科学院计算机网络信息中心作为中科院“十二五”科技数据资源整合与共享工程项目的总承担单位，秉承“统筹规划，整合集成，公开共享，服务科研”原则，践行由硬件建设向环境构建、工程化项目向持续化发展的重要转变，构建云服务模式，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。其中，基础设施即服务（IaaS）提供数据的云存储、云计算、云灾备、云归档等服务；数据即服务（DaaS）支持自助云端数据建库管理，推动数据在云端汇聚，以通用接口实现数据云共享；软件即服务（SaaS）则基于云环境支持各类数据应用软件的发布、运行和共享。数据工程整体架构设计如**图1**所示。

中科院“十二五”数据工程项目共设置数据存储与管理云服务环境、科学数据整合与共享服务、海量科学数据分析与应用示范3个相互紧密联系又独立实施的子项目。

“数据存储与管理云服务环境”子项目面向科学活动大数据的管理和应用需

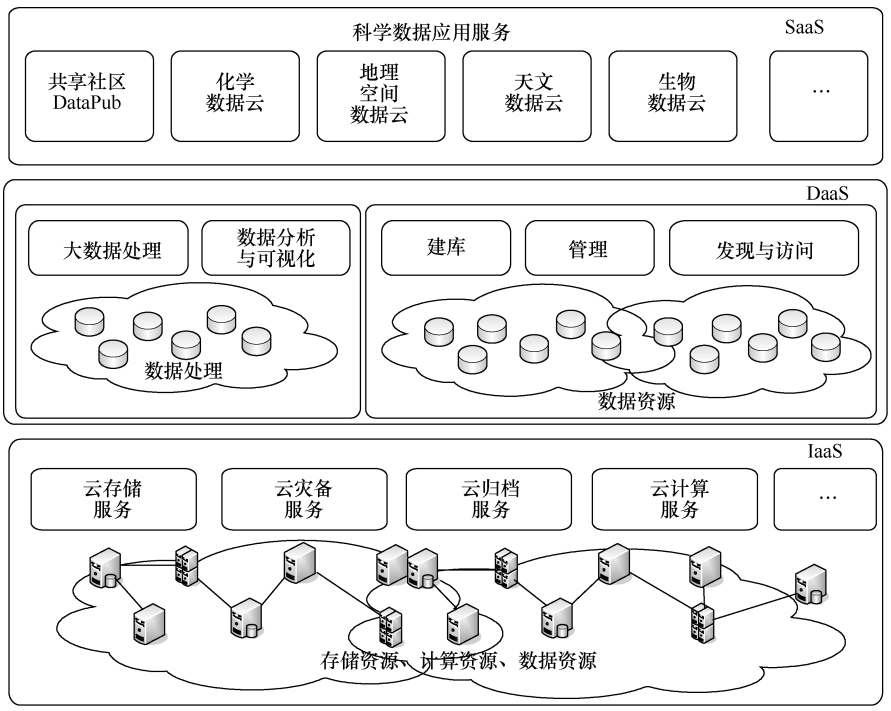


图1 “十二五”数据工程整体架构

求，建设具有海量存储与处理能力的科学数据基础设施。在“十二五”末期形成50 PB容量的院级存储与服务环境，布局全院、直达各所，实现存储设施的虚拟化统一管理；与先进网络设施互通，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；为海量科学数据管理和共享提供运行支撑环境，为“十二五”创新活动提供存储设施保障。

“科学数据整合与共享服务”子项目面向学科发展和科研应用，通过整体规划和设计，在全院公开优选资源，重点整合一批具有优势地位的学科领域主题数据库，示范整合一批研究所数据资源整体集成的专题数据库，滚动支持一批长期积累和能够共享的专业数据库；形成科学数据共享服务监控和效果评估管理机制和支撑系统，深化数据资源的集成整合服务，推动全院的数据整合、归档、汇聚和发布共享，整合资源量达到500 TB，并作为数据云服务的核心内容，深入融合在数据基础设施，并提供广泛的公共数据服务。

“海量科学数据分析与应用示范”子项目依托海量存储为核心的数据基础设施，立足全院海量数据资源和科技文献信息，加强数据挖掘分析与可视化系统，加强数据与文献服务的集成化服务系统，实现科学数据与科技文献语义关联服务示范，实现数据可视化交互分析平台，充实数据基础设施的基本服务，深化先进数据应用，形成具有特色的面向公共支撑服务的海量数据分析与应用环境。

3 “十二五”数据工程整体建设情况

中科院“十二五”数据工程以数据资产为核心，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，也是大数

据科研成果服务于社会应用的示范平台。数据存储与管理云服务环境、科学数据整合与共享服务以及海量科学数据分析与应用示范3个子项目得以顺利实施并取得了丰硕的成果。

（1）建成了存储容量达52 PB的分布式的海量存储环境，支撑全院重要数据资产的容灾备份、长期保存、共享服务与增值应用

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。截至项目结束，中科院数据云存储环境运行服务总容量达52 PB，云存储规模达8 PB，共拥有物理服务器约300台，虚拟机5 000多台的计算服务能力；数据归档总容量达38 PB，拥有归档能力大于20 TB/天、在线磁盘阵列容量达到2 PB、近线磁带库存储容量达到30 PB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”的高等级的灾备服务。

同时，该项目研发部署了可视化数据管理与发布工具VisualDB、数据交换与共享云平台 DataPub、科学数据服务效果评估工具等，形成了面向科学数据领域的云存储和云计算服务，为科研人员提供了稳定易用的数据库建设工具、自组织的科研社区建设工具，形成了支持科学数据库绩效评估和科学数据出版的能力；支持了空间科学先导专项、卫星遥感、微生物等领域的数据存储，在35个院内单位部署了VisualDB，为北京市地方税务局等地方单位提供了数据容灾服务；科学数据引用规范、数据溯源表达模型获得国家标准立项，《中国科学数据》获得我国首批网络连续型出版物试点（CN11-6035/N）。

（2）面向科技数据资源的持续发展与应用，通过重点库与专业库建设，基本形成后评估模式的科学数据长期共享服务环境和管理机制

“科学数据整合与共享服务”子项目面向中科院科学研究活动的需求，强化科学数据资源的整合与集成，基本形成后评估模式的科学数据长期共享服务环境和管理机制。截至项目结束，科学数据库重点完成了资源学科领域、土壤学科领域、动物学科领域、植物学科领域、材料学科领域等13个领域重点库，紫金山天文台、昆明植物研究所、南海海洋研究所、南京地理与湖泊研究所等7个所级重点库的整合建设，完成了大气科学、黄土高原水土保持、中国湿地与黑土生态、中国“金钉子”等20个专业库的持续建设与服务，数据资源内容广泛涉及地学、生物、物理、化学、材料、空间、天文、海洋、能源、信息等学科领域，数据量、数据质量、学科覆盖范围均得到大幅度提升。

根据科学数据库统一监控与统计分析，58家建库单位共建完成40个数据库，整合可共享的资源量达655 TB。特别是依托科学数据库共建单位建立了面向全院的数据咨询服务体系，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了良好的效果，积极推进了典型的数据应用。“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。

（3）基于科学数据与文献关联服务应用示范和海量科学数据分析可视化关键技术研究与应用示范的成功探索，为用户提供强大的科学数据与文献互联、丰富的可视化展示等功能

“科学数据与科技文献集成服务关键技术研究与应用”示范课题的主要目标是采用近期热点研究的开放关联的理念，进行实践探索，将科学文献与科学数据有

效关联，为科学信息的获取和传播探索新途径。充分利用关联数据的关联机制，通过关联映射模板及唯一标识符实现数据层不同类型资源描述框架（resource description framework，RDF）资源间的关联，同时采用了数据挖掘机制，通过术语共现分析等手段开展关联路径分析，使得文献与科学数据间的内在关联和外部关联均得到充分展示。

“海量科学数据分析可视化关键技术研究与应用”示范课题面向可视化应用开发，基于模型驱动理论，设计并采用了可视化应用模型——DVDL，利用模块化、层次化描述的可视化描述语言，可对组成可视化的各个部分进行不同抽象层次上的描述。其研究成果已经应用到中科院寒区旱区环境与工程研究所、地理科学与资源研究所等研究单位的科研活动，同时在伊利集团、北京市疾病预防控制中心、北京市地方税务局等企事业单位进行了实际的使用，产生了相应的社会效益和经济价值。

3个子项目的顺利实施，推动了中科院科技数据基础资源、海量存储与处理基础设施、数据集成与应用先进环境的建设与服务，形成以海量科学数据为核心的系列“海·云”服务，成为科技云的重要支柱之一。“十二五”末期，着力提升科技数据战略管理和支撑服务能力，为中科院乃至国家科技发展提供强大和持续的数据基础设施。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相关技术以及持续推动科学数据云发展打下了坚实的基础。

4 技术创新

4.1 面临的问题与挑战

科学数据的大规模主要体现在数据量大、分布广泛、结构多样等方面，而科学数据的服务要求快速有效，这对数据服务体系的设计和实现提出了问题与挑战。

（1）如何快速整合资源

科学数据的格式类型和存储形式多种多样，结构化、半结构化、非结构化数据同时存在，如气象数据、地学数据等都有其独特的数据结构和存储方式；同时，目前的科学数据作为重要资源，往往掌握在各科研单位手中，分布极为分散，这种分布式的存储形式对数据快速有效的组织和整合形成了障碍。如何将这些多源异构的数据管理和集成起来并提供统一快速的服务，是数据服务体系需要解决的重要问题之一。

（2）如何针对资源提供高质量的数据服务

数据服务体系的最终目标是为科研人员提供高质量的数据服务，这对服务的组织形式、交付方式都提出了较高的要求。一个好的服务体系设计需要对服务模式、交互方式等有深入的研究和分析。由于可持续管理是服务体系长期运行和有效服务的关键环节，与政策等密切相关，同时也对技术架构提出了要求，通过技术设计促进服务体系的有效管理并形成激励。

4.2 技术整体架构

针对大规模数据资源分散存储与统一服务的总体需求，结合上述体系框架设计的问题与挑战的分析，在整个体系框架设计中，采用分层设计的模式，以满足不同层次管理与服务的需求。整体框架分层结构如**图2**所示。

整个服务体系框架共分3层，自底向上分别是自治管理层、整合管理层、集成服务层。其中，自治管理层重点实现分布式数据资源自治管理与服务，完成数据资源的本地化集成注册、服务封装及发布管

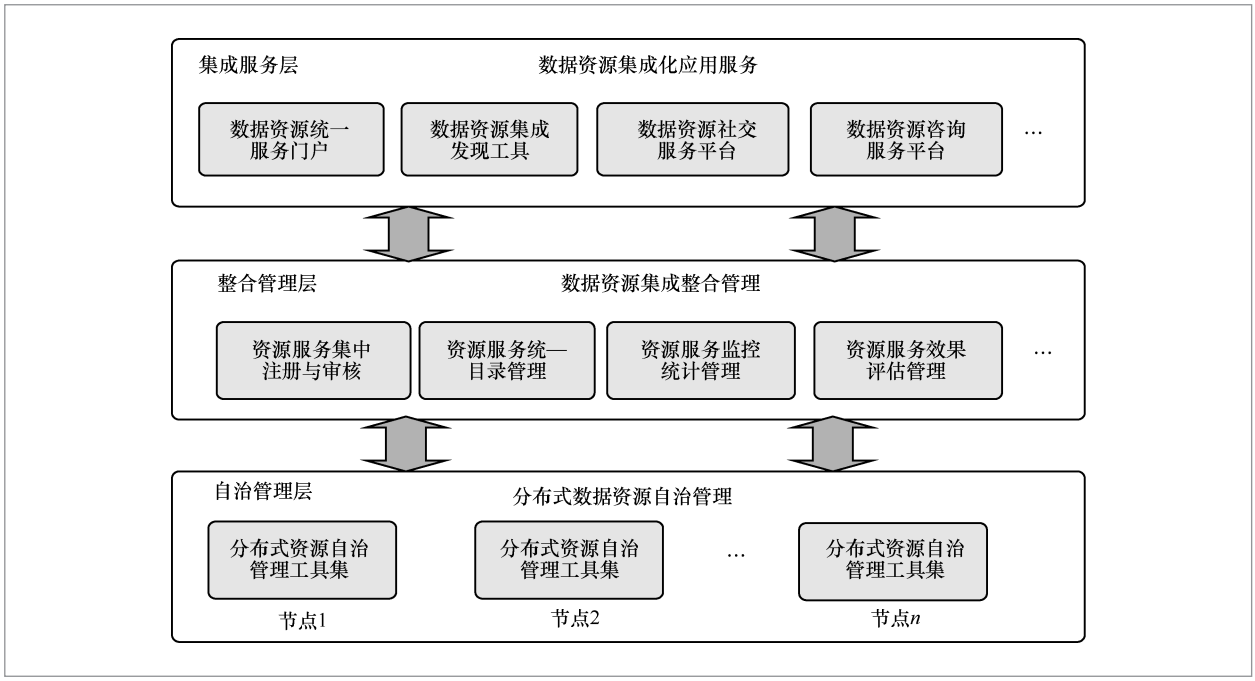


图 2 大规模分布式科学数据管理与服务体系分层框架

理；整合管理层重点实现数据资源与服务的集中注册、审核与发布管理，进而形成统一的资源服务目录，同时实现对数据资源与服务的监控、统计和评估管理，为分布式数据资源与服务的稳定、优质服务提供支撑和保证；集成服务层是整个体系的对外服务门户，该层重点实现数据资源的目录、发现、访问、获取等公共服务，同时面向最终用户提供以数据资源为中心的集成、交流、共享、咨询方面的服务系统。

4.3 典型的工具和软件

下面从系统分层框架自治管理层、整合管理层和集成服务层各选一个典型的工具软件进行重点介绍。

（1）自主建库与共享服务工具集VisualDB 3.0[①]

为解决分布式数据的在线获取速度慢、不稳定，数据分散化严重，不好组织和整理，不能提供稳定API等问题，研发自主建库与发布工具集VisualDB3.0，为本地独立、自治的数据库提供可视化、可配置的数据管理与发布功能。VisualDB3.0是一个帮助数据管理者管理和发布关系型数据库和文件系统的工具，帮助应用研发人员快速开发面向数据应用的研发框架，是一套帮助数据应用低成本集成异构数据源的解决方案。通过多途径在线建库、自助式数据管理、定制化数据发布，科学数据库能够为e-Science应用提供组织良好、质量有保障、可稳定访问的数据，目前支持MySQL、SQL Server、Oracle等多种关系数据库以及文件数据类型等，还可以通过数据访问的接口直接访问数据，用VDB Server来完成数据迁移。

（2）科学数据服务监控与评估系统[②]

针对分布式数据库网站的多类多指标采集的评估需求，整个评估体系建设根据评估指标的特点，进行分类集中采集和注册汇交，形成了相关支撑系统层。数据监控与访问统计系统完成中断运行时间、访问

① http://www.vdbspace.cn/
② http://sees.csdb.cn/

人次、在线下载量等定量指标的采集，资源量在线统计系统记录数、数据容量、数据更新频率、访问接口正常率等定量指标的采集，咨询服务系统完成服务响应率定量指标的采集，资源与服务注册系统完成元数据、服务案例、知识产权的相关定性指标的集中注册与汇交。全部定量指标的采集均可通过自动模式获取。

(3) 中科院数据云门户③

中科院数据云门户从降低设计复杂度、提升可维护性和可扩展性的角度出发，软件应用整体设计采用MVC架构，对模型维护、数据展示、请求与响应进行了分层处理。控制层使用Spring MVC作为框架，Spring MVC能够很好地与Spring框架集成，并支持REST风格，框架稳定、性能优异；使用Spring作为容器来管理服务层的控制逻辑处理，能够很好地解耦层次之间的调用关系；数据处理层采用Spring Data框架，能够适应于大多数数据库；为了能够集成科学数据资源与服务注册系统的查询功能，采用Axis2作为Web Service技术，Axis2能够与Spring框架集成，使用广泛。同时针对新闻采编需求，定制开发了采编系统，提供页面静态化、广告、访问统计等功能。中科院数据云门户提供各项应用监控信息和数据统计功能，系统采用定时轮询和多任务处理方式，采集各项应用服务的监控和统计数据，并使用可视化方式进行展示。

5 服务创新

5.1 数据服务科研新模式

“十二五”期间中科院数据云形成了以基础设施云服务、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，并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开发为示范，进行了数据服务科研新模式的思考和探索。

大数据资源库的开发和服务探索了一种多源异构大数据融合管理与服务的新方式。大数据时代的科学研究面临的关键需求和技术挑战包括海量复杂数据低成本高可靠存储、高效管理以及快速分析与服务等。面向海量多源异构数据管理问题，本研究突破了其中关键的技术问题：面向异质科学数据的一致化管理技术、多源数据流程化汇聚与加工技术以及基于大数据资源库的在线数据服务技术等。通过构建适合大规模科研数据的存储集群和管理系统，最终实现对结构化数据、非结构化数据和半结构化数据的融合管理。项目以生物学领域数据为示范，设计开发大数据资源库服务平台，通过生物学领域数据对外数据服务接口，为中科院微生物研究所、广州生物医药与健康研究院的有关应用提供数据服务。

科学数据出版提出了一种科学数据发布和引用的新模式。科学数据出版是科研人员与科研机构按照统一规范的质量管理和控制机制，主要利用互联网及其他方式公开发布其通过观察、实验、计算分析等科研过程所产生的原始数据（raw data），或通过对已有数据进行系统化的收集、整理和再加工，形成数据及数据产品（data product）的出版行为④。科学数据出版通过对科学数据相关利益者权益的梳理，试图化解数据开放共享的诸多问题，帮助使用者便捷地发现、获取、理解和再分析利用数据，并可在科研论文及相关科研成果中引用数据。2015年8月，中科院计算机网

③ http://www.csdb.cn

④ http://www.csdata.org/paperview?id=9

络信息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⑤。

⑤ http://csdata.org/

数据众包服务示范了一种基于互联网的数据产品加工与服务新方式。众包作为互联网一种崭新的生产组织形式，通过高效调用分散的人力资源实现海量数据快速精准分析的方法[2]。随着互联网和计算机技术的发展，众包在各行各业得到了广泛应用，其中也包括科学数据采集和处理领域。中科院计算机网络信息中心开发的地理空间数据云（GSCloud）是一个基于云计算技术的海量地学数据资源以及数据处理模型服务的平台，自2007年开始向公众提供服务以来，至今已经积累了海量数据资源，形成了完善的基础设施和专业的服务团队，累计注册用户13万人次⑥。2015年5月，GSCloud发布了第一个数据处理众包任务，并在整个工作过程中，建立了整套的流程框架，包括需求明确与任务划分、任务发布与分配、数据处理、质量控制、结果集成和报酬发放等部分。目前GSCloud共发布了36个任务，吸引和积累了大量专业人才，人才库中的专业人员达到1 100余位。

⑥ http://www.gscloud.cn/

此外，数据工程项目还开发了科学数据共享社区——DataPub。作为一个融合社交网络理念的数据共享和交流平台，DataPub构建了多层次共享服务框架，实现了数据的有效流通与便捷访问，为用户提供了多途径、高质量的交互服务。其中，数据共享层满足数据发布、检索、访问等以数据为中心的共享服务；数据社交层围绕用户开展数据社交、定制化需求等个性服务；数据融合层设计不同领域数据的集中管理与融合。通过该共享平台，用户能够进行数据发布，让更多人知晓和获取数据，发挥数据价值；可以查找和获取DataPub上的数据，或将数据需求提交到平台；还能够促进朋友圈交互、数据社区交互，实现全方位的数据交流与互动。

5.2 数据服务典型应用案例

“十二五”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。资源学科领域基础科学数据以俄罗斯、蒙古等“一带一路”国家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，发挥了为“一带一路”建设决策和国家治理提供长期的科技战略咨询的作用。多民族语言资源数据库推动了“一带一路”区域文化与科技交流，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。

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。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。南海海洋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。“面向政府决策的湖泊水环境治理决策与预警”数据专题服务，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，在太湖和巢湖的蓝藻调查、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。

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。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。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。中微子实验自正式取数以来，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，此为世界最高精度。基于中国植物物种信息数据库编著的《中国植物志》出版后，中国科学院昆明植物研究所率先提出了“iFlora研究计划”，拟基于《中国植物志》的研究成果，打破传统意义上的纸本和单一产品《中国植物志》界限，实现植物物种多样性研究的标准化、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的需求。“iFlora”研究计划的提出，开辟了后植物分类学的新时代。

科学大数据孕育科研方法的新范式。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。从海量数据中解析其蕴含的新模式，科学大数据正带来科研方法论的新范式，如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，高能天体物理数据库已经成为我国空间天文科学体系中的重要组成部分。《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了科学依据，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。

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展和改革委员会、国家食品药品监督管理总局、北京市地方税务局等30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、高校以及相关企业的大批人员参加。

6 未来展望

通过“十二五”整体建设和深化应用，“科技数据资源整合与共享工程”部署形成了共建共享的海量存储基础设施运行服务环境，协作推进政策、环境和管理契合科学数据共享良性发展的新模式。通过推行数据云服务先进的发展理念和有效的运行机制，有力地引导和整合科学数据基础性工作，将科学数据战略机遇转化，实现为数据云服务，抢占数据密集型科学发现的制高点和前沿阵地。

“十三五”期间，在国务院《促进大数

据发展行动纲要》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。同时，在深入大数据驱动的科研创新应用的基础上，聚焦科学大数据基础性理论问题研究和相关关键技术的突破，引领国内科学大数据的发展。

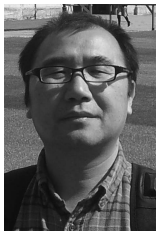
中科院数据云将考虑构建可以承载大数据资源、支撑大数据分析、推动大数据应用的可扩展平台环境，具有支持PB级大数据应用处理能力；支持实现一批大数据应用技术的研发部署，为科学大数据相关研究、测试和应用提供基础条件，为大数据应用技术研发、培训和示范服务等提供实验环境；营造和制定实施科学大数据的相关环境、机制和标准规范，为协调推动全院大数据发展行动、夯实大数据应用与研发环境提供基本指导和规范，良好促进科学大数据的建设发展；最终力争实现立足中科院、面向科技界，形成共享开放、服务创新的国家级科技数据中心。

参考文献:

[1] 郭华东. 大数据、大科学、大发现[J]. 中国科学院院刊, 2014, 29(4):500-506.
GUO H D. Bigdata, big science, big discovery[J]. Bulletin of Chinese Academy of Sciences, 2014, 29(4):500-506.

[2] BARRINGTON L, GHOSH S, GREENE M, et al. Crowdsourcing earthquake damage assessment using remote sensing imagery[J]. Annals of Geophysics, 2011, 54(6): 680-687.

作者简介

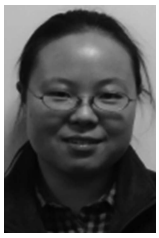


黎建辉（1973-），男，博士，中国科学院计算机网络信息中心研究员、博士生导师，大数据技术与应用发展部主任，CODATA中国委员会秘书长，主要研究方向为大数据管理、大数据分析与处理。

周园春（1975-），男，博士，中国科学院计算机网络信息中心研究员、博士生导师，主要研究方向为大数据分析与挖掘。

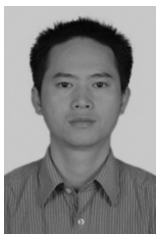
胡良霖（1973-），男，中国科学院计算机网络信息中心高级工程师，主要研究方向为数据库技术与标准规范、数据质量与数据服务。

刘峰(1974-),男,中国科学院计算机网络信息中心高级工程师,主要研究方向为科学数据管理与服务体系构建。



朱艳华(1982-),女,中国科学院计算机网络信息中心高级工程师,主要研究方向为数据库技术与标准规范、数据应用服务。

沈志宏(1977-),男,博士,中国科学院计算机网络信息中心高级工程师,主要研究方向为科学数据管理与共享、关联数据、大数据管理。



吴章生(1980-),男,中国科学院计算机网络信息中心工程师,主要研究方向为地图学与地理信息、大数据技术与应用。

张杨(1982-),男,中国科学院计算机网络信息中心工程师,主要研究方向为数据库技术与标准规范、数据应用服务。

收稿日期: 2016-10-08
基金项目: 中国科学院“十二五”信息化基金资助项目(No.XXH12504)
Foundation Item: Special Project of Informatization of Chinese Academy of Sciences in “the Twelfth Five-Year Plan” (No.XXH12504)