

西部数据中心的数据服务实践

李红星 王 建 南卓铜 王亮绪 吴立宗 潘小多 李 新

(中国科学院寒区旱区环境与工程研究所, 甘肃兰州 730000)

摘 要: 科学数据是人类科学活动成果的积累, 对于科技创新、经济发展具有重要的战略意义, 科学数据共享是社会发展的必然需求。国家自然科学基金委员会“中国西部环境与生态科学数据中心”(简称西部数据中心)面向“中国西部环境与生态科学计划”及广大科研工作者展开科学数据服务, 并已取得显著的服务成效。数据中心采取的多层次的共享政策及主动、专业、专职的服务小组是数据共享服务顺利进行的重要保障。数据发布方、数据作者与用户三方交流机制的建立及面向科学研究的数据服务是进一步发展的方向。同时, 数据共享机制不健全、国家层面的共享环境没有形成是目前阻碍数据共享服务发展的瓶颈。

关键词: 中国西部环境与生态科学数据中心; 数据服务; 数据共享; 科学数据

中图分类号: TP3

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2010.03.004

Data Service of Environmental and Ecological Science Data Center for Western China

Li Hongxing, Wang Jian, Nan Zhuotong, Wang Liangxu, Wu Lizong, Pan Xiaoduo, Li Xin

(Cold and Arid Regions Environmental and Engineering Research Institute, CAS, Lanzhou 730000)

Abstract: Scientific data is the accumulation of the achievements in scientific research, which plays an indispensable role in technological innovation and economic development in a long run. Scientific data sharing is an inevitable demand of social development. "Environment and Ecological Science Data Center for Western China", sponsored by the National Natural Science Foundation of China (NSFC), provides scientific data service for the NSFC'S "Environment and Ecological Science Program" and the majority of researchers, and dramatic accomplishments had been achieved. Multilevel sharing policy and active, professional and qualified staffs are the guarantee of the successful data sharing. The further striving goals in the next step are to build the communication network among data publishers, data owners and data users, and to develop data service facing concrete scientific research. Meanwhile, the lack of perfect data sharing mechanism and circumstances in the national level is the bottleneck to impede the development of data sharing service.

Keywords: environment and ecological science data center for western China, data service, data sharing, scientific data

1 引 言

科学数据是人类科技活动积累的宝贵财

富, 是科学创新的源泉^[1]。系统地进行科学数据集成与共享将极大地促进地球系统科学研究的可持续发展, 为地球系统科学长远发展与重点突破提供强有力的支撑^[2]。国外现有的科学数据共享

作者简介: 李红星(1981—), 女, 工程师, 主要研究方向: 数据共享和地理信息系统。

基金项目: “中国西部环境与生态科学研究计划”重点项目“中国西部环境与生态科学数据中心”(90502010); 中国科学院西部行动计划(二期)项目“黑河流域遥感—地面观测同步试验与综合模拟平台建设”(KZCX2-XB2-09)。

收稿日期: 2010年5月6日。

机制已经比较成熟，专门从事数据收集、交换、服务和共享的国际组织相继成立，欧美发达国家也将数据共享提到战略高度^[3-9]。美国更是将“完全与开放”的数据共享政策作为一项基本国策：联邦政府资助的科学数据，即公共性、基础性的国有数据，必须在没有歧视的基础上以不超过复制和发行成本的费用无限制地使用^[10]；私有科学数据则采取自由竞争政策和市场化共享机制。科研人员及社会各阶层人员通过完善的国家共享数据库群及科学数据共享服务网即可获取需要的数据及详尽的数据说明文档。

我国现有的科学数据资源已经较为丰富，但科学数据共享与知识产权保护存在着共享意识薄弱、产权界定不清、利益分配不当和宏观引导不够等诸多问题^[11]，科学数据共享程度仍然比较低，资源浪费、重复建设的现象普遍存在。因此，要打破科学数据壁垒，实现科学数据共享，在推进数据共享的政策、法规日臻完善与成熟的同时需要建立一个专门的数据收集、整理与发布的科学数据共享服务平台，提供完整的数据服务功能，提高数据的利用效率。

国家自然科学基金委员会“中国西部环境与生态科学数据中心”，简称“西部数据中心”，是国家自然科学基金委员会地学部“中国西部环境与生态科学计划”（以下简称“西部计划”）资助下建立的数据共享平台，由中国科学院寒区旱区环境与工程研究所和中国科学院地理科学与资源研究所承建和运行，其任务是收集、管理“西部计划”项目数据产出，并面向“西部计划”各项目 and 国内外科研人员提供科学数据服务。西部数据中心的定位首先是直接服务于“西部计划”的科学目标，建成地区特色鲜明，信息高度综合，突出数据集成，同时又能带动整个地球表层科学研究的数据中心^[12]。其服务领域主要集中在干旱区水文水资源研究、流域科学研究、西部生态环境变化研究、冰冻圈研究等。自运行以来，西部数据中心已经为数百个科研项目和数千名用户提供了数据服务，为环境与生态研究提供了良好的数据共享平台。

2 数据形式

鉴于国内目前的数据共享现状，还无法面向所有用户实现“完全与开放”的数据共享。西部数据中心力求在保护数据安全和数据作者权益的基础上最大限度地开展科学数据的有序共享。基于此，服务形式采取在线和离线相结合的方式：对于遥感影像及其各类产品、较小比例尺专题图等可公开数据实行在线共享；国家相关法律法规规定的涉密数据及数据生产单位有共享范围限制的数据则通过离线方式共享，用户须通过必要的审核并签署共享协议方可获取数据，协议对数据版权、使用限制、保密义务、引用规范等细则均有详尽描述（见西部数据中心在线使用条款和免责声明 <http://westdc.westgis.ac.cn/about/terms>）。“西部计划”所有项目则拥有最高级别共享权限。这种对数据和用户分级的多层次共享方式，一方面保护了数据安全及数据作者的权益，另一方面可以保障科学数据合理流通。用户在西部数据中心网站搜索到需要的数据，在线数据注册登录后即可下载，数据通过 FTP 发布，以用户名和密码控制安全性。对于离线数据，首先在线填写数据共享协议，包括申请人姓名、单位、联系方式等个人信息和数据用途，然后将协议打印签字并邮寄或传真至数据中心，经审核通过后即可获取数据下载链接或以光盘形式邮寄（图 1）。

为了及时响应用户的离线申请，提供优质、流畅的数据服务，西部数据中心成立了专门的数据服务小组全面负责数据共享事宜，并设有专人全职处理相关工作。此外，由于数据用户来自多

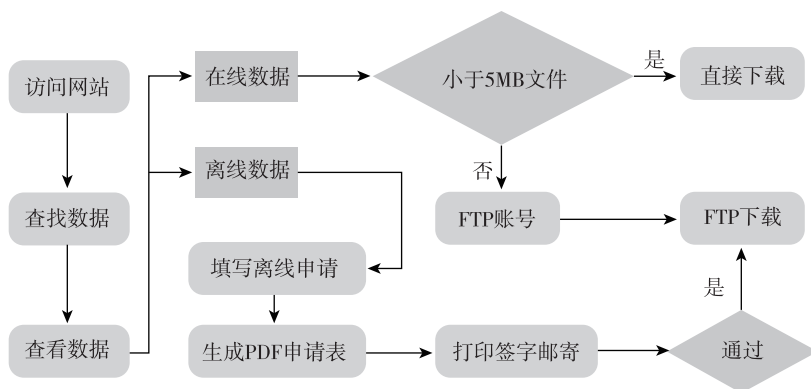


图 1 数据申请下载流程

个行业部门,具有不同的学科背景,对数据的认知与使用能力参差不齐,元数据和数据说明文档中的信息有时不能满足所有用户的需求,研究者需要通过邮件或电话的方式咨询数据使用过程中存在的各种问题,设立固定的数据服务小组可以让研究者找到明确的咨询对象。同时,数据服务人员通过与用户之间的交流,可以了解数据、平台、服务等环节存在的问题,对于数据管理和数据质量控制等具有重要意义。首先,及时了解用户在下载数据过程中遇到的问题,便于进一步完善数据管理模式,提高数据共享的流畅性;其次,数据使用情况的反馈让数据质量控制能够有的放矢,提高数据质量控制的效率;再次,深入的交流能够更多地获取用户的数据需求信息,对于数据中心整理、生产数据具有指导意义。

此外,针对“西部计划”项目采取了主动推送的数据服务方式。数据服务小组于2006—2007年分别对北京、西安、成都等14个分布有西部项目的城市进行了直接有效的登门拜访活动,为各“西部计划”项目提供数据中心已有数据的同时深入调查了各项目的数据需求,并就数据共享和数据服务中可能存在的一系列问题进行了广泛的交流,为进一步搞好数据共享工作、明确数据中心发展方向收集到了第一手资料,一些启动较早的项目汇交了当时已经产出的部分数据。

3 相关规范与数据汇交

目前,国内保障国家安全的国家数据机密管理已经比较成熟,建立了有效的安全保障系统,对于测绘、气象、水文等数据均有相关的立法。但是针对数据共享还没有专门的法规和条例可循。为了进行规范、安全的科学数据共享服务,西部数据中心制定了《科学数据与资料集成和共享管理条例》,对于数据收集、管理、共享和保护的方式和范围等均作了详细说明。同时,制定《数据服务人员工作细则》对服务的质量和时效进行控制。此外,还制定了数据共享协议,明确规定了数据引用标注规范、数据严禁用于商业用途及第三方转让以及数据保密等相关条

款,以保护数据生产者的著作权及数据安全。通过这些相关条例的制定和遵守,很大程度上规范、优化了数据服务工作,同时对数据用户也给予必要的约束,达到保护数据引用机制和著作权的目的。

集成各“西部计划”项目的产出数据并建立西部环境与生态数据库是西部数据中心的科学目标之一。为了实现对接数据的有效管理,“西部数据中心”制定了一套科学的数据处理流程,首先对接数据进行整理、分类、格式转换等规范化处理,同时编写标准化的数据文档,并多次反馈给数据作者修改、定稿。此外,在确定数据发布时间和共享方式时,必须与数据提供者进行密切协商,制定可行的数据共享政策。

西部数据中心运行以来,对“西部计划”项目数据进行了持续汇收。截至2009年底,全部66个“西部计划”项目中共有28个汇交了数据。西部数据中心为所有汇交数据编写了元数据,并在数据中心网站公开发布。此外,还将汇交数据单独或分类集成为特色数据集,并统一编写数据说明文档。但相对于66个西部计划项目来说,汇交数据的项目十分有限,仍有大部分项目未提交数据。这不仅是“西部计划”实施中存在的问题,而且是国内整个科技界普遍存在的问题。主要是由于国内数据共享的大环境没有形成,国有科学数据共享管理机制不健全,公益化和产业化两种机制的划分不十分清楚,其中最主要的问题是对科学数据的产权划分不清楚^[14]。很多由国家投资生产的科学数据,仅在数据生产单位甚至更小的团体内流通,很难获取。在获取数据非常困难的情况下,必然激发对自我数据的保护,如此便陷入获取困难导致数据保护,数据保护又使获取更加困难的恶性循环。此外,科学家提交的数据得不到安全、合理的共享,数据整合部门不能

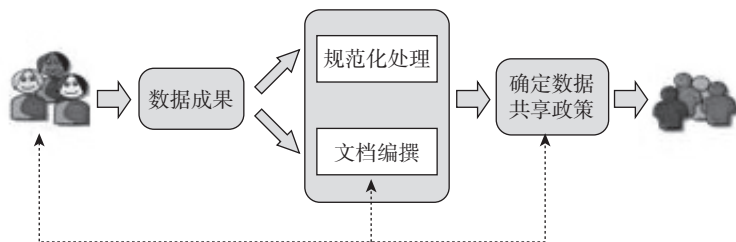


图2 西部计划汇交数据处理流程

对数据进行有效整理、发布，数据用户不能正确引用数据等都是阻碍数据汇交的因素。

4 服务成果

截至 2009 年底，数据服务总量达 6.2 TB，已向 100 多家科研单位的科研人员提供离线数据服务达 388 人次，离线数据申请量达 1.86 TB，答复数据和技术咨询 1500 多人次；网站注册用户达 2667 人，在线数据集下载 6550 次，下载总量 4.34 TB。表 1 为西部数据中心从 2005 年底到 2009 年主要运行情况。西部数据中心网站共得到 6 万多用户 30 余万次访问，并且访问量呈逐年大幅提升趋势；数据申请 2006—2007 年以离线为主，2008 年随着在线共享平台的正式投入使用，大量数据可以直接从在线共享平台获取，在线下载量迅速上升，并大幅超越离线申请量，总数据申请量和数据用户亦逐年增加。

在数据共享服务的同时，数据中心还在不断收集、生产新数据，补充、更新已有数据。截至 2009 年底，数据中心共拥有共享数据 5.1 TB，范围涉及生态、水文、冰雪、气象、人口经济及各类背景数据，数据类型有专题数据、观测数据、遥感数据、科学试验数据、模型数据等。数据中心服务对象的学科领域主要集中在干旱区水循环、水资源管理、生态环境评价、生态服务功能与安全、土地覆盖变化、气候变化、陆面过程研究、冰冻圈研究等方面。申请、下载的热门数据主要以各类 NDVI 产品，Landsat TM/ETM 影像，流域科学数据（黑河、石羊河等），全国各类专题数据（沙漠、植被、土壤、土地利用等），长时间序列雪深、冻融数据，被动微波亮温数据，土地覆盖综合数据，冰川编目数据等为主。

在数据共享平台愈加完善与成熟，用户不断增长，数据服务量快速攀升的同时，数据服务不能满足科研人员数据需求的问题也越来越显现出来。一方面是因为数据中心目前数据类目还不够齐全、部分数据时空范围不足或数据精度不能满足要求；另一方面是因为一些已有的数据出于对数据生产者权益的保护无法对所有用户展开共享。可以看出，单纯的技术提升已不能从根本上

推动数据共享的发展，这对创建良好共享环境，加强数据整合提出了迫切的要求。

5 经验总结

5.1 元数据与数据文档

元数据和数据文档是数据的重要组成部分和有效补充。元数据提供数据摘要、格式、时空范围、学科分类、下载与引用等基本说明，同时能提高数据检索效率。完整规范的元数据对于用户从大量数据中快速找到目的数据有很大帮助。目前，西部数据中心已经完成了对所有发布数据元数据的编写与修改。但对于来自不同背景的用户，在数据使用中仅有从元数据获取的数据基本信息往往是不够的，用户需要了解数据采集、处理、制备的详细资料，同时希望得到数据质量评价和使用限制等高级信息，因此完备的数据说明文档是非常必要的。数据中心自产数据的文档内容已经比较详尽，但收集的其他数据由于信息不易获取，高质量文档的制备还有一定困难，这在很大程度上影响了服务的效果，要解决该问题还需要与数据生产者建立更多的协作。

5.2 广泛的交流平台

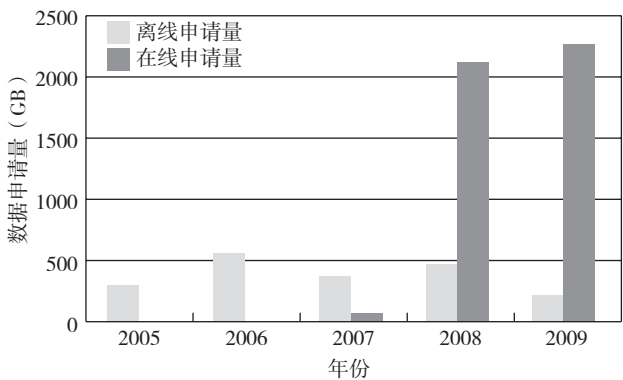
由于地球科学数据涉及领域广泛，专业性极强，数据中心不可能配备各个学科的专业人才，对于数据质量的控制通常也只能满足技术层面的需求，用户反馈的问题不一定能全部解答。因此，数据中心需要成为联系数据生产者与用户的桥梁，提供专门的交流平台，建立生产者与用户及用户间的广泛沟通。数据中心不仅共享数据，更要促进知识、创新、经验的共享。

5.3 面向科学研究的数据服务

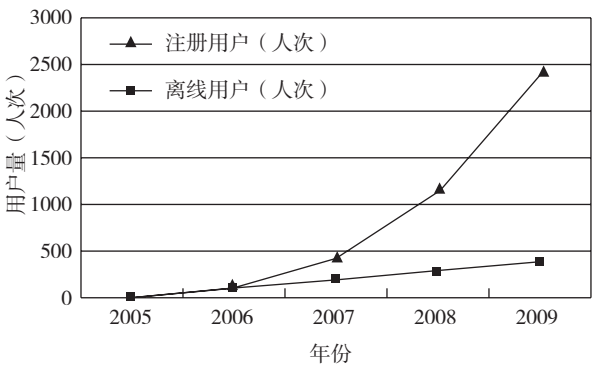
首先，数据服务应该打破被动等待的服务模式，针对重大的科学计划主动提供共享服务。其次，服务不应仅仅停留在简单提供已有数据的层面，而应该加强与用户的深入交流，掌握明确的科学问题与数据需求，并以此为导向，从数据的制备、数据质量和数据组织等各方面满足科学研究的需要。以此为导向的数据库建设已经超越了一般意义上的“数据库建设”，已经不能从数据的外部和技术层面来理解、管理和提供数据，必须

表 1 西部数据中心 2005-2009 年主要运行情况

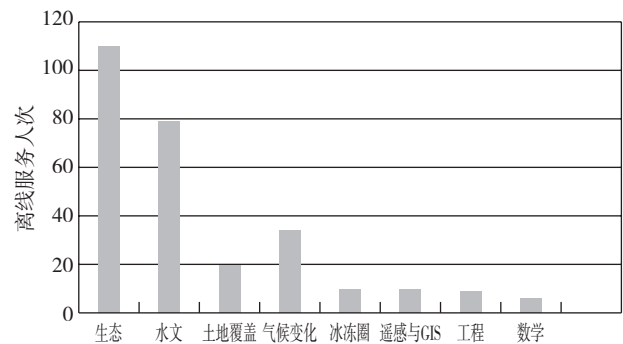
运行指标	2005	2006	2007	2008	2009	合计
数据总量(GB)	556	1629	1973	2171	5243	—
网站访问用户(个)	298(11-12月)	7,124	11,253	18,085	29,669	66,429
离线申请次数	19	76	94	83	116	388
离线申请量(GB)	285.83	559.54	368.58	474.44	228.63	1907.2
在线下载量(GB)	0	0	61.39	2114.07	2271.92	4447.38
总申请量(GB)	285.83	559.54	429.97	2588.51	2490.55	6364.4



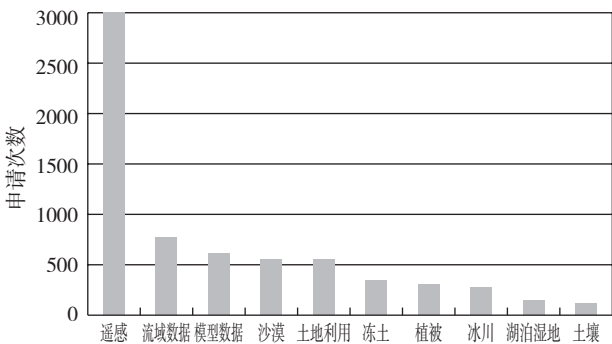
a.西部数据中心2005-2009逐年数据申请量



b.西部数据中心2005-2009年数据用户增长情况



c.按照研究学科分类的离线申请量分布情况



d.按照数据类型的申请量分布情况

图 3 中国西部环境与生态科学数据中心数据服务统计

在充分认知数据的基础上，从数据的内部来重新组织和提供数据，并可能产生新的数据。只有面向科学研究的数据服务，才能在数据中心实现真正意义上的学科交叉和项目合作，才能建立公共数据平台，这种需求迫使数据中心必须建立一支“工程师+科学家”式的队伍。

5.4 科学数据共享环境

我国科学数据共享与知识产权保护存在的诸多问题导致数据资源分散在不同部门与个人手中，数据整合与共享难度大，数据中心难以实现对各类科学数据的集成，很多数据需求无法满

足，数据服务也因此受到严重的限制。所以，要建立公平、良好的数据共享环境，必须从国家层面上调整科学数据共享管理机制，明确产权，区分公益性和产业化两种不同的运行机制，并颁布正式的政策法规进行法律保障。与此同时，健全数据发布、引用等相关机制，充分保护数据生产者的知识产权。对于后者数据中心可以起到很好的推动作用：数据中心在管理时必须明确数据的来源，准确地标明数据的生产者或贡献者，数据的生产者和贡献者不应因数据存储单位的变更而变更。数据中心应尽可能地为每个数据提供一个

建议引用方式, 用户在使用数据时, 数据中心也必须明确要求用户正确地引用该数据。经过一段时间后, 只有当数据使用者普遍自觉地尊重别人的数据成果, 并正确地使用别人的数据成果, 数据生产者才有更大的积极性去整理和共享数据。

6 结 论

科学数据的广泛共享是社会发展的必然需要, 是科技创新、经济发展的必要保障。数据服务是数据共享的有效途径, 优质的服务能够提高数据共享的效率。国内数据共享机制不健全, 共享环境还没有形成, 是阻碍数据服务和共享发展的重要瓶颈, 大量极具价值的数据无法流通, 数据集成困难重重, 技术方面的进步和数据共享机构单独的努力已经不能从根本上推动数据服务的提升, 国内数据共享事业急需进行国家层面的共享意识培养及法制保障。其次, 数据服务需要更多关注科学需求, 主动提供数据支持, 甚至参与到科学研究中去, 不但提供数据, 还可以帮助科研项目管理、发布新的产出数据。最后, 促进数据服务方、数据生产者和用户的三方交流机制也是数据服务今后工作的重点, 对于深入了解需求, 解答疑问、促进创新都具有很大的意义, 是对共享的扩大和深化。

参考文献

- [1] Sun Jiulin, Shi Huizhong. The Construction of Data Sharing Network of Earth System Science in China[J]. China Basic Science, 2003(1):76-81. (in Chinese)
[孙九林, 施慧中. 中国地球系统科学数据共享服务网的构建[J]. 中国基础科学, 2003(1):76-81.]
- [2] Li Xin, Nan Zhuotong, Wu Lizong, et al. Environmental and Ecological Science Data Center for West China: Integration and Sharing of Environmental and Ecological Data[J]. Advances in Earth Science, 2008, 23(6):628-637. (in Chinese)
[李新, 南卓铜, 吴立宗, 等. 中国西部环境与生态科学数据中心: 面向西部环境与生态科学的数据集成与共享[J]. 地球科学进展, 2008, 23(6): 628-637.]
- [3] Li Juan, Liu Dehong, Jiang Hong. Research on International Scientific Data Sharing[J]. Library Development, 2009(2):19-25. (in Chinese)
[李娟, 刘德洪, 江洪. 国际科学数据共享现状研究[J]. 图书馆建设, 2009(2):19-25.]
- [4] NSF Cyber Infrastructure Council. NSF Psyche Infrastructure Vision for 21st Century discovery [R/OL]. [2008-07-18]. <http://www.nsf.gov/news/special-reports/cyber/index.jsp>.
- [5] Science and Innovation Investment Framework 2004 - 2014 :next steps [EB/OL].[2008-07-18]. <http://www.hm-treasury.gov.uk/media/7/8/bud06-science-332v1.pdf>.
- [6] National e - Science Centre [EB/OL].[2008 - 07 - 18]. <http://www.nesc.ac.uk>.
- [7] The Bucharest Declaration, Bucharest Pan—European Conference in Preparation of the World Summit on the Information Society: Towards an Information Society: Principles, Strategy and Priorities for Action [R]. Bucharest, 2002, 11.
- [8] Developing the UK Pse-infrastructure for Science and Innovation [EB/OL]. [2008-07-18]. <http://www.nesc.ac.uk/documents/OSI/index.html>.
- [9] Jiang Hong, Zhong Yongheng. Study of the International Scientific Data Sharing[J]. Journal of Modern Information, 2008(11):56-58. (in Chinese)
[江洪, 钟永恒. 国际科学数据共享研究[J]. 现代情报, 2008(11):56-58.]
- [10] Policy Statements on Data Management for Global Change Research[EB/OL]. [2008-04-22]. <http://www.gcric.org/USGCRP/DataPolicy.html>.
- [11] Liu Chuang. Scientific Data Sharing Mechanism in US and Its Implications for China[J]. China Basic Science, 2003(1):34-39. (in Chinese)
[刘闯. 美国国有科学数据共享管理机制及对我国的启示[J]. 中国基础科学, 2003(1): 34-39.]