中国科技资源导刊 ISSN 1674-1544 CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
2010 年 9 月 第 42 卷第 5 期 30-36 ISSN 1674-1544 Vol.42 No.5 30-36, Sept. 2010

西部数据中心数据集成和共享的回顾与展望

王亮绪[1] 南卓铜[1] 吴立宗[1] 冉有华[1] 李红星[1] 潘小多[1] 祝忠明[2] 李 新[1] 丁永建[1]
(1. 中国科学院寒区旱区环境与工程研究所，甘肃兰州 730000；
2. 中国科学院资源环境科学信息中心，甘肃兰州 730000)

摘 要：中国西部环境与生态科学数据中心承担西部计划项目数据产出的收集、管理、集成，并面向西部环境与生态科学的各个领域提供科学数据服务，形成了从数据收集、规范化整理、集成挖掘到数据服务的体制，建成了功能丰富的数据共享网站系统，集成了一批西部环境与生态乃至整个中国陆域地球表层科学方面的关键数据集，为西部计划等项目及科研团体与个人提供了持续的数据服务。文章总结了西部数据中心的总体框架、数据集成和数据服务，并探讨了西部数据中心持续发展的方法。

关键词：科学数据中心；数据共享；中国西部环境与生态科学数据中心；数据集成
中图分类号：G **文献标识码**：A **DOI**：10.3772/j.issn.1674-1544.2010.05.006

Environmental and Ecological Science Data Center for West China: Review and Outlook

Wang Liangxu[1], Wu Lizong[1], Nan Zhuotong[1], Ran Youhua[1], Li Hongxing[1], Pan Xiaoduo[1], Zhu Zhongming[2], Li Xin[1], Ding Yongjian[1]
(1. Cold and Arid Regions Environmental and Engineering Research Institute, CAS, Lanzhou 730000;
2. Scientific Information Center for Resources and Environment, CAS, Lanzhou 730000)

Abstract: The Environmental and Ecological Science Data Center for West China (WESTDC), sponsored by the National Natural Science Foundation of China, is aiming to collect, manage, integrate and disseminate environmental and ecological data for western China, as well as build a multi-functional data-sharing system. This paper summarizes the overall framework , the data integration and the data services of WESTDC. The outlook of WESTDC is also described to be a scientific data center for cold and arid regions in China.

Keywords: scientific data center, data sharing, Environmental & Ecological Science Data Center for West China, data integration

1 背 景

国家自然科学基金委员会于 2001 年启动了针对西部开发战略的重大科学研究计划“中国西部环境与生态科学研究计划”(以下简称“西部计划”)。西部计划的总体目标是逐步实现对中国西部环境变化过程和机理的整体性深入认识，为解

第一作者简介：王亮绪(1976-)，男，工程师，研究方向：科学数据共享、陆面数据同化系统。

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)和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 9 月 14 日。

决西部环境和可持续发展中的重大科学问题提供更加充足的依据[1]。西部计划启动以来，已经逐步形成了围绕西部环境和生态领域重大科学问题开展交叉协同研究的平台。自2005年起，在西部计划执行后期，西部计划的综合集成工作启动。数据集成是集成工作的重要内容之一，因此在国家自然科学基金委员会资助下于2005年启动了“中国西部环境与生态科学数据中心”（以下简称“西部数据中心”，http://westdc.westgis.ac.cn）。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，建成地域特色鲜明、信息高度综合、突出数据集成，同时又能够带动整个地球表层科学研究的地球科学数据中心[2]。西部数据中心的主要任务包括：（1）整理和规范现有的西部环境与生态数据，为西部计划提供数据服务；（2）集成西部计划的科学数据与研究成果，促进西部计划的数据共享；（3）通过数据与文献的关联管理，逐步建立西部环境与生态数字图书馆。同时探索数据共享和再分析的有效机制，成为西部计划促进学科交叉、实现项目间合作与交流、进行西部环境生态科学集成研究的重要支撑平台。

西部数据中心自建立以来，生产、整理和集成了大量科学数据，建成了一个功能丰富的地球科学数据共享平台，并在数据共享服务方面取得显著成效。本文归纳总结了西部数据中心建设过程中的组织结构、数据生产集成、共享政策、数据平台、数据服务等，展望了下一步西部数据中心的发展目标，以期更多的科学家和公众了解、支持和使用西部数据中心，提升科学数据共享氛围。

2 总体框架

西部数据中心的总体框架设计以西部计划和地球系统科学发展的科学需求为导向，以便于西部计划各学科交叉研究为准则，为西部计划成果的综合集成及地球科学发展的长远目标服务。在图1中，数据共享平台提供数据共享功能，合作交流平台提供信息交互功能，知识积累平台通过文献库提供知识管理功能，数据科学平台提供数据集成研究功能。这4个平台联合起来组成西部数据中心面向用户的网站系统。

西部数据中心是在国家自然科学基金委员会地球科学部的直接领导下，依托中国科学院寒区旱区环境与工程研究所建设运行，数据集成组负责数据的整理、生产和集成，撰写元数据和数据文档，平台建设组负责搭建数据共享所需的各种功能，数据服务组负责西部数据中心的对外交流、数据咨询以及离线服务等，数据指导与协调委员会负责制定数据共享和保密的各种政策并实施于西部数据中心（图2）。数据集成组由许多不同研究方向的专业研究组组成，这些专业组本身

图1 西部数据中心总体框架图

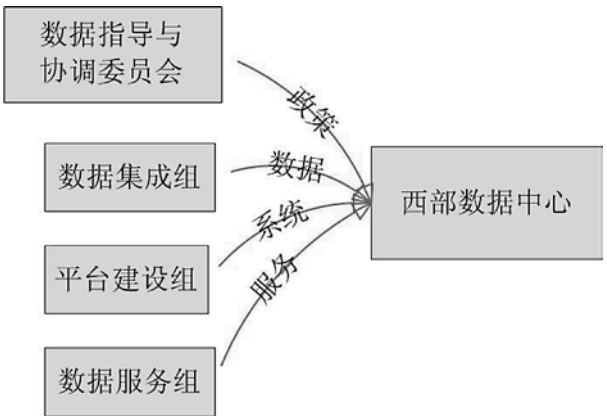

图2 西部数据中心组织结构图

从事相关领域的实际研究工作，了解研究前沿和实际需求，与环境与生态研究领域联系密切，同时与自身研究目标不冲突，甚至可促进自身研究任务的完成。

2.1　数据共享平台

数据共享平台定位于数据集的管理发布并向用户提供数据信息和数据服务，并基于先进的网络技术，建立分布式的西部数据中心数据共享系统，充分发挥地域和学科优势，形成层次分明、力量集中、特色鲜明的数据中心，促进有利于科学研究的数据共享氛围。西部数据中心以元数据为核心搭建数据共享平台[3]，管理后台实现对元数据的创建、编辑和同步，在前台通过元数据库实现数据的导航、查找、搜索和下载等功能，同时提供各种 Web Services 开放标准服务功能。具体功能包括：支持西部数据中心特色数据集的快速共享，支持离线数据的在线申请和管理，基于唯一标识符实现了数据的固定连接；基于元数据库设计了灵活的数据快速导航系统，实现了空间数据的不同导航模式，包括分类、关键词、数据集序列、时间和空间导航方式，实现了空间数据的快速搜索、浏览，可以让用户快速地找到所需的数据。数据共享平台的 Web 页面如图 3 所示。

图 3　数据共享平台网站页面

2.2　知识积累平台

知识积累平台的实现主要通过综合运用描述、链接和集成等多种方式，建成支持面向西部环境与生态科学研究的综合信息利用平台。其设计目标为：（1）汇集西部计划研究项目所产生的知识产出；（2）建立有关西部环境与生态研究的专题数据库，逐步发展和形成西部环境与生态科学文献平台；（3）对环境与生态科学领域的学术信息进行搜索、发现、聚集和再组织，建立环境与生态科学开放知识库系统。知识积累平台基于开源软件 DSpace 作为快速开发和构建知识积累平台的基础，总体功能和服务设计遵循 OAIS 参考模型，具备对知识内容进行采集、描述、存储、保藏、发布、浏览和利用等多方面功能和服务[4]。通过人工和自动方式，完成了中国西部环境和生态科学研究计划、环境与生态期刊资源、环境与生态科学信息网关、环境与生态开放知识资源、西部环境与生态科学文献库、黑河流域研究、地球科学数据导航等知识专题的组织建设（图 4），并已

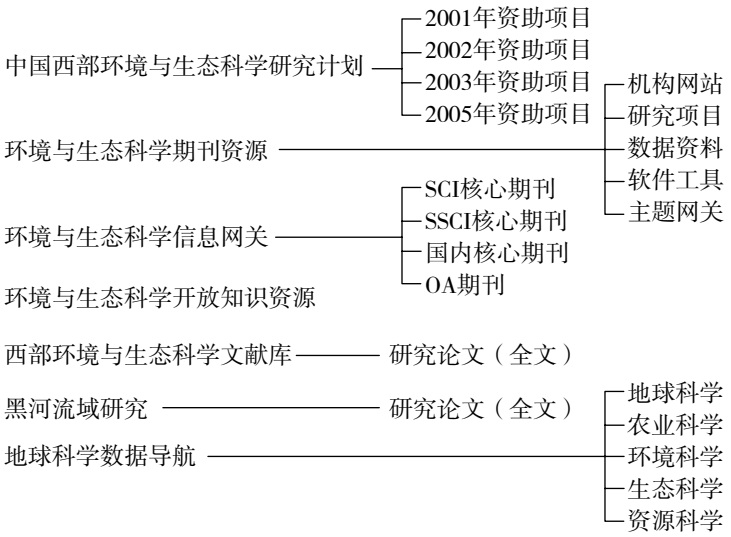

图 4　知识积累平台总体内容组织结构

采集积累了 6 万多条知识资源条目。

2.3 合作交流平台

合作交流平台的主要功能是提供互动交流的技术工具，包括：（1）开通内容丰富的网上论坛，通过学术论坛、邮件列表、即时通讯等方式，提供一个针对数据以及科学问题的自由讨论空间，实现科学家间的密切交流和合作，促进学科交叉。（2）汇集和提供西部计划的基本信息以及与之相关的信息，发布项目动态、学术活动等。

合作和交流平台提供数据用户之间、数据用户与数据提供者之间的沟通机制，在一个虚拟的平台上将不同学科对某一问题感兴趣的科研人员组织起来，而不受空间地域的物理限制。通过相同或不同学科基于网络的自由讨论，可以碰撞产生一些优秀的想法。

2.4 数据科学平台

数据科学研究是数据中心的高级阶段，也是对集成化目标的重要诠释，更是体现“环境和生态领域整体创新活动的公共平台”的重要特征。主要包括：（1）对基础数据进行加工处理，通过与西部计划内科学家的密切合作，形成若干种可被广泛使用的模型数据集和具有可靠质量的再分析资料等标准数据集。（2）开展数据融合、数据集成、数据同化等数据科学的前沿研究。

3 数据、元数据和数据文档

数据是科学数据中心运行的基础和核心，数据建设要面向数据集成和应用，数据、元数据和数据文档是数据建设的基本要素，其中元数据和数据文档是对数据的描述和说明，是数据的必要补充信息。西部数据中心整理、生产和集成了大量的满足中国西部环境与生态科学所需的科学数据。依照西部计划的科学目标，西部数据中心首先实现了对现有数据资源的系统整合，避免了数据库重复建设。在此基础上广泛收集和整理与生态环境研究相关的国内外数据资源，为西部计划提供基础数据支持；对数据中心的数据资源进行再次遴选，挑选出具有专业特色的质量较高的数据集或数据集系列，并撰写了详细的数据文档，制备出可供方便使用的特色科学数据集。

3.1 数据

根据西部计划项目的特点及综合集成工作的需要，集成的数据包括：（1）背景数据。主要是指反映西部环境与生态现状与变化的基础性空间数据，包括：遥感数据、气象和水文资料、西部地区背景空间数据、社会经济人文数据。（2）流域综合数据，包括黑河流域数据库、石洋河流域数据库、天山乌鲁木齐河流域基础数据库。（3）环境与生态观测数据。该数据来自于单个项目的零散观测资料，往往不够系统，除已经数字化的资料外，部分数据也散布于研究报告和论文中。它们能够支持对特定地域的过程研究，也能够支持特定模型的模拟，但数据的结构性较差，如果不经过规范化和标准化处理，就很难在信息系统中进行管理，也难以被其他科学家所使用。而对这些数据的处理又显然是与科学问题密切联系的，需要和西部计划内的科学家之间密切配合。（4）大型综合试验数据，包括 1989–1993 年在黑河中游开展的地气相互作用试验（HEIFE）、全球能量水循环之亚洲季风青藏高原试验（GAME–Tibet）、全球能量与水循环协调观测计划（CEOP）在蒙古的观测数据、黑河综合遥感联合试验数据[5]。（5）模型数据集。是依据地球系统科学对高分辨率观测和模型模拟的要求，从观测资料经过集成处理和质量控制而得出的一般性的可用于环境、生态和地球表面过程数值模拟的单点数据集和网格化数据集。模型数据集包括驱动数据、验证数据、参数集以及同化资料。详细的数据列表请见西部数据中心网站。

3.2 元数据

元数据是描述数据的数据，对于地球空间数据，包括数据摘要、格式、覆盖区域、下载和引用等信息。元数据不仅能够对数据进行基本说明，而且能够搭起不同数据中心间交互的桥梁。科学数据中心在进行数据共享时都会通过元数据来扩展数据共享的功能。西部数据中心元数据采用 ISO 19115 元数据国际标准，若其他的数据中心采用不同的元数据标准，可以通过元数据工具进行标准的转换。西部数据中心对每条数据都撰写了元数据，并在元数据中着重突出了数据作者的知识产权信息以及数据贡献者的信息，提供了

中英文数据引用格式，以促进科学数据的引用和出版。

3.3　数据文档

科学数据不同于一般的数据。元数据所提供的数据描述通常是不够完备的。用户希望了解数据背景、数据采集、数据处理等细节。例如从遥感数据得到的生物物理参数集，用户希望能够从大气校正到参数反演算法的数据处理细节上都能给予详尽说明。此外，数据中心往往只能在技术上对数据进行质量控制，科学数据需要经过类似科学论文的同行评议，才能更好地被科学家所接受，因此，必须提供完整的数据文档，用户才能对数据做出取舍和判断。同时，在地球表层科学越来越成为实验科学的今天，数据文档也是科学实验可重复性的重要原始证据。

西部数据中心的数据文档包括[6]：（1）数据处理过程。包括元信息、有关数据产生的科学目标等背景信息、数据的采集方法和取样方法、仪器的相关说明、数据处理的方法和详细过程。（2）数据使用说明。包括数据格式的读取，数据使用建议等。（3）数据来源及引用信息。为了突出对数据产权的保护，特别在数据文档中提供明确的数据来源和数据引用的相关信息。西部数据中心制作的数据文档列表请见文献 [6]。

4　数据服务

西部数据中心在运行以来，依据自己的数据服务理念，制订了相关的数据规范和数据汇交办法，实行了主动的推送式数据服务和离线数据服务，并取得了一系列的数据服务成效[7]。

4.1　数据服务理念

数据服务是指数据中心如何利用本身的数据资源为科学研究服务，这种服务不应该仅仅是被动式的，还应该包括主动式服务。西部数据中心的数据服务理念有两个层次：（1）面向科学研究的数据服务。指以科学问题为导向，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必须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。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，这就要求数据中心必须建立起一支“工程师 + 科学家”式的队伍。（2）培养数据共享的公共意识。要建立公共数据研究平台，数据共享是基础。这种数据共享并不仅仅是指数据中心向科学家提供数据，而是要在科学家群体之间建立一种自愿的、公平的数据共享和使用的良好学术环境。建立这种学术环境的核心需要建立数据成果评价机制、知识产权保护机制和数据使用引用机制。第一种机制的建立不属于数据中心的范畴，而后两种机制的建立，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。经过较长时间的实践，只有当数据使用者能够自觉地尊重他人的数据成果，并正确地使用他人的数据成果，数据生产者才能有更大的积极性去整理和共享数据。

4.2　数据共享策略

科学研究工作向来注重开放，对于从事科学数据共享服务的人来说，更需要一个积极、开放的态度。只有有了这种开放的态度，在制定具体的数据共享规范的时候，才能尽量地扩大科学数据服务的对象，取消各种不必要的限制，对地球科学研究提供最大程度的数据支持。而阻碍数据应用的主要障碍是数据中心共享政策不明朗、数据共享权限划分复杂、收费与不收费界限不清晰。西部数据中心采用“完全与开放”（Full & Open）的数据共享政策。所有的科学家或研究项目都有权获得数据中心的数据以及数据处理过程等技术文档，西部数据中心以无差别的、不高于复制和邮寄成本的形式向所有数据用户提供数据。西部数据中心对注册用户不进行权限限制，仅对用户的数据下载使用情况进行记录。“完全与开放”的数据共享理念是西部数据中心各项措施得以落实的思想基础。

4.3　推送式数据服务

数据服务是数据中心与用户联系的桥梁，高

质量、高效率的服务将极大地推动数据中心数据的使用。传统的数据中心一般都进行被动式的数据服务，只有当数据用户有要求的时候才提供数据，并且不注重数据的对外宣传，数据与用户之间缺乏联系的桥梁。因此有必要采取推送式的数据服务。西部数据中心主要采取以下措施来实现推送式的数据服务：（1）简化数据申请程序。保证大多数数据在线下载。（2）建立固定的数据服务小组。固定的数据服务小组主要负责数据门户信息系统所无法解决的数据用户特殊需求，为用户解决数据方面的疑惑，出版数据通讯，对用户数据需求情况进行调查等。（3）出版数据通讯。数据通讯是国际上重大合作项目普遍采用的一种形式，在信息交流和推动数据应用方面的作用非常突出，初期以数据内容通报为主，逐渐过渡到数据内容与数据应用和集成方法并重。

4.4 数据服务成效

截至 2010 年 8 月，西部数据中心可共享数据为 5.1TB，对外数据服务总量达 9.2TB，其中在线下载量约为 7TB ，已向 100 多家科研单位的科研人员的 18 个 973 项目、3 个 863 项目、33 项西部计划项目、44 个自然基金项目、18 个国家科技支撑项目、13 个科学院项目、100 个普通项目、86 篇硕博论文及 122 个其他科研项目提供了数据服务，并检索到在 96 篇期刊文章和硕（博）士论文中标注了数据来源于西部数据中心。

5 讨 论

5.1 数据资源的共建共享

数据中心的可持续服务，只靠一批项目、几家单位是绝对不行的。为了保持数据资源的持续发展，西部数据中心在自主开发新数据的同时，通过合作引进国内外相关领域的科学数据集，同时欢迎相关领域的科学家在西部数据中心发表自己的数据成果，吸引国内外相关的重大观测研究项目通过西部数据中心发布数据。西部数据中心将在做好知识产权保护的基础上按照数据作者的要求做好数据的共享和服务，以更大程度地放大科学家及项目成果的价值。同时，西部数据中心也将联络其他地学数据中心，互通数据资源，并通过 Web Services 技术实现不同数据中心的互通共享。

5.2 推动数据发表机制

目前，我国科学数据的发表机制尚未形成，科研人员的数据发表意识不强。西部数据中心针对科学数据共享的现状开展了初步的探讨，从完善元数据、数据文档特别是数据的引用信息开始，保证数据用户能够方便获得数据的引用信息，并保证数据作者的知识产权不受侵权。同时，采用数字对象唯一标识符（DOI）对每条数据进行注册，以保证数据的唯一性和权威性，并通过 DOI 来追踪和分析数据的使用情况。这些措施将有利地推动数据发表机制的形成。

5.3 提升数据质量，打造科学数据中心品牌

科学数据质量的高低，直接决定了数据能否使用和共享，因此数据中心需要通过技术手段提升科学数据质量，包括从数据自身、元数据及数据文档等 3 个方面进行数据的质量控制和评价。西部数据中心下一步将通过类似期刊文章的同行评议手段对数据中心进行质量控制和评价。数据中心品牌效应的形成对于数据中心的长期有效运行非常重要。但如何让数据用户使用数据放心、让科技界承认数据中心数据可靠等，是一个长期的过程，西部数据中心还需要进一步努力，争取打造科学数据中心品牌。

6 结论与展望

国家自然科学基金委员会在基础研究领域和重大科学计划的组织上有着很高的信誉和号召力。在目前国内大力推动数据共享但现状还难以令人满意的情况下，国家自然科学基金委员会可以在整合地学数据资源方面发挥其他部门和机构所难以比拟的优势，大力倡导完全与开放的共享原则，通过稳定的支持和承建单位的不懈努力，把西部数据中心建成一个国家级的有重要影响的地学数据中心，为地学基础研究和西部地区社会经济可持续发展研究作出贡献。西部数据中心面向西部环境与生态科学的各个领域提供科学数据服务，形成了从数据收集、规范化整理、集成挖掘到数据服务的体制，建成了功能优化的数据共

享网站系统，集成了一批西部环境与生态乃至整个中国陆域地球表层科学方面的关键数据集，为西部计划等项目及科研团体与个人提供了持续的数据服务。西部数据中心将继续在数据集成的广度和深度方面努力，并与其他数据中心合作，争取建立以环境与生态领域的科学问题为导向，紧密服务在寒区旱区开展的各类科学计划的特色科学数据中心。

参考文献

[1] Leng Shuying, Li Xiubin, Cheng Guodong, et al. The Progress of Studies on the Environmental Change and Ecological Issues in Western China[J]. Science Foundation in China, 2005(5):262−267. (in Chinese)
〔冷疏影 , 李秀彬 , 程国栋 , 等 . 中国西部环境和生态科学重大研究计划阶段性进展及深入研究的问题 [J]. 中国科学基金 , 2005(5): 262−267. 〕

[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] Wang Liangxu, Wu Lizong, Nan Zhuotong, et al. Application of Open Source Technologies in Geoscientific Data Centers[J]. China Science & Technology Resources Review, 2010, 42(3):17−23, 35. (in Chinese)
〔王亮绪 , 吴立宗 , 南卓铜 , 等 . 开源技术在地球科学数据中心中的应用 [J]. 中国科技资源导刊 , 2010, 42(3): 17−23, 35. 〕

[4] Zhu Zhongming, Ma Jianxia, Chang Ning, et al. A DSpace−based Sharing Environmental and Ecological Knowledge Space[J]. Library and Information Service, 2007, 51(4): 71−74, 108. (in Chinese)
〔祝忠明 , 马建霞 , 常宁 , 等 . SEEKSpace——基于 DSpace 的环境与生态科学知识积累平台 [J]. 图书情报工作 , 2007, 51(4): 71−74,108. 〕

[5] Li X, Li X, Li Z, et al. Watershed Allied Telemetry Experimental Research[J]. Journal of Geophysical Research, 2009, 114: 19.

[6] Pan Xiaoduo, Li Xin, Nan Zhuotong, et al. Research on Data Description Document [J]. China Science & Technology Resources Review, 2010, 42(3):30−35. (in Chinese)
〔潘小多 , 李新 , 南卓铜 , 等 . 科学数据文档的研究 [J]. 中国科技资源导刊 , 2010, 42(3): 30−35. 〕

[7] Li Hongxing, Wang Jian, Nan Zhuotong, et al. Data Service of Environmental and Ecological Science Data Center for Western China [J]. China Science & Technology Resources Review, 2010, 42(3):24−29. (in Chinese)
〔李红星 , 王建 , 南卓铜 , 等 . 西部数据中心的数据服务实践 [J]. 中国科技资源导刊 , 2010, 42(3): 24−29. 〕

（上接第 21 页）

[22] Wang Liangxu, Wu Lizong, Nan Zhuotong, et al. Application of Open Source Technologies in Geoscientific Data Centers[J]. China Science &Technology Resources Review, 2010, 42(3): 17−23, 35.(in Chinese)
〔王亮绪 , 吴立宗 , 南卓铜 , 等 . 开源技术在地球科学数据中心中的应用 [J]. 中国科技资源导刊 , 2010, 42(3): 17−23, 35. 〕

[23] Pan Xiaoduo, Li Xin, Nan Zhuotong, et al. Research on Data Description Document[J]. China Science & Technology Resources Review, 2010, 42(3): 30−35.(in Chinese)
〔潘小多 , 李新 , 南卓铜 , 等 . 科学数据文档的研究 [J]. 中国科技资源导刊 , 2010, 42(3): 30−35. 〕

[24] Fayyad U, Piatetskyshapiro G, Smyth P. The KDD Process for Extracting Useful Knowledge From Volumes of Data[J]. Communications of the ACM, 1996, 39(11): 27−34.

[25] Ran Y H, Li X, Lu L. Evaluation of Four Remote Sensing Based Land Cover Products Over China[J]. International Journal of Remote Sensing, 2010, 31(2): 391−401.