文章编号:1000-0240(2010)05-0970-06

中国西部环境与生态科学数据中心在线共享平台的设计与实现

南卓铜[1], 李 新[1], 王亮绪[1], 丁永建[1], 祝忠明[2], 吴立宗[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: 国家自然科学基金委员会的"中国西部环境与生态科学数据中心"致力于中国西部环境与生态科学数据的共享. 在介绍西部数据中心整体结构的基础上, 对西部数据中心的在线数据共享平台的技术实现方案、开发模式和数据库结构, 元数据标准、数据知识关联、分类和导航等关键技术实现进行了介绍. 西部数据中心共享平台本着方便查询、详细数据说明、易于管理、着眼未来和突出服务等原则, 充分利用最新的 Web 和 GIS 技术在政策许可范围内实现数据共享. 最后, 讨论了数据中心之间集成、模型服务等问题.

关键词: 数据中心; 在线数据共享平台; 中国西部; 元数据; 数据科学

中图分类号: TP311.13 **文献标识码**: A

0 引言

科学数据是开展科学研究活动, 尤其是定量研究的基础[1], 科研人员通过分析科学数据加深对现象和规律的理解. 数据情况也往往决定着建模方式和模型简化程度, 也影响研究开展的深度和广度[2]. 环境和生态领域的科学数据来自室内外试验、野外观测以及航天、卫星遥感等途径[3]. 对于广袤的中国西部, 观测站点稀少, 环境和生态数据十分有限, 因此就显得更为珍贵. 2000 年开始的国家自然科学基金委员会(以下简称基金委)中国西部环境与生态科学研究计划(以下简称西部计划), 资助了一系列针对西部环境和生态的科研项目[4]. 正是认识到西部数据的现状, 2006 年基金委启动了"中国西部环境和生态科学数据中心"(以下简称西部数据中心)项目, 旨在基金委西部计划项目内实现数据共享, 减少数据重复投资, 并为整个西部的地球表面过程研究提供数据服务[5]. 建立数据中心是目前国际上普遍采用的数据共享方式, 通过集中数据存储, 建立数据服务机制, 实现数据共享.

在环境和生态领域, 国际上已经建立了一些大型的数据中心, 并取得宝贵的数据共享经验[6]. 世界数据中心(WDC, http://www.ngdc.noaa.gov/wdc/)在 12 个国家, 建立了地球物理、环境、人文等 52 个中心, 极大推动了国际间的科学数据共享[7]. 美国宇航局(NASA)建立的对地观测系统数据信息系统(EOSDIS)[8]是美国主要卫星数据的共享系统, 负责数据的存储、管理、发布以及数据产品生成、数据工具开发和分析处理, 数据量每天以数 TB 增长. www.geodata.gov 是美国政府实现电子政务的重要数据门户, 依托 GIS 技术实现了地图等空间数据的发布、浏览和搜索一站式服务[9]. 美国 Oklahoma 州 Mesonet 气象资料共享网 Agweather 通过互联网发布 40 个气候观测站的实时监测和预报数据(http://agweather.mesonet.org/), 是国际上实时监测数据共享的成功例子之一[10]. 事实上, 各个政府部门、学术组织、大学研究所、公司、行业都纷纷建立了各自的数据中心. 同时一些国际数据协会如 CODATA(http://www.codata.org/)也在推动国际间的数据共享和提高各国的科技数据管理和使用技术.

在我国, 近些年的数据共享也取得长足的进

收稿日期: 2010-01-18; **修订日期**: 2010-03-30

基金项目: 国家自然科学基金重大研究计划项目"西部生态与环境科学数据中心"(90502010)资助

作者简介: 南卓铜(1977—), 男, 浙江乐清人, 副研究员, 2003 年在中国科学院寒区旱区环境与工程研究所获博士学位, 现主要从事空间决策支持系统、空间建模环境、数据中心和地理信息系统研究和开发. E-mail: nztong@lzb.ac.cn

展．比如，科技部组织的科学数据共享工程旨在整合分散的科学数据资源，创建社会化的共享服务体系[11]，受科学数据共享工程的支持，中国气象局的中国气象科学数据共享网加大气象数据的共享力度(http：//cdc. cma. gov. cn/)[12]；中国科学院启动了科学数据库及其应用项目，同样为了整合分散的科学数据库群[13]．然而无论在共享程度、影响力、数据科学的研究深度和技术的先进性方面，我国的数据共享离国际水平仍有一些距离．

西部数据中心(http：//westdc. westgis. ac. cn)依托国家基金委，于 2007 年正式提供服务，截止 2008 年 9 月 26 日，在线数据量达 842GB，向 100 余家教育和研究单位提供了数据服务．有关西部数据中心总体情况和当前数据产品请参考李新等[5]文章．本文主要介绍西部数据中心的在线数据共享平台的设计和实现，重点在西部数据中心特色功能的考虑和实现上，最后讨论了数据中心之间的集成等问题．

1　在线数据共享平台

1.1　西部数据中心整体结构

西部数据中心包括数据共享平台、知识积累平台、合作交流平台和数据科学平台(图 1)[5]．数据共享平台包括在线数据共享和离线数据服务，在线数据共享子系统通过 Web 页面提供了数据发布、检索和下载．考虑到特定数据的共享限制(比如国家数据政策的限制以及与数据所有方签署的共享协议的限制等)，西部数据中心设置了专人负责离线数据申请服务，用户需要签署纸质数据申请书以明确数据责任才能获取数据．离线数据服务在目前的数据共享制度下仍是不可取代的．西部数据中心坚持“全面和开放”(Full & Open)[14]数据共享政策，在政策允许范围内，最大限度实现数据共享．知识

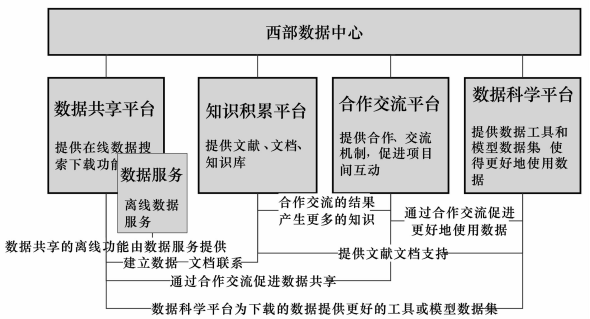

图 1　西部数据中心总体结构

Fig. 1　The overall architecture of Ecological and Environmental Science Data Center for the West China (WestDC)

积累平台依托中国科学院资源环境科学信息中心的文献库，链接数据和文献、文档和数据说明，是数据知识化、应对长期数据存储导致数据知识丢失[15]的一个尝试．合作交流平台通过网络论坛、邮件列表等手段提供了合作、交流机制，促进用户和数据中心之间的交流．数据科学平台提供了数据工具和模型数据集，开展数据挖掘研究．如图 1 所示，各平台间相互联系，共同实现西部数据中心的总目标．

1.2　设计原则

数据中心的最根本功能是数据的共享．数据共享平台设计遵从以下原则：

(1) 快捷的数据查询．用户通过数据共享平台能方便快捷地找到所需的数据，并以最低的成本、最快捷的方式下载到所需的数据．

(2) 详细的数据说明．数据有足够详细的元数据，以记录数据的内容、质量、处理方法、表示方式、空间参考等信息．同时提供数据说明文档以补充元数据对数据背景知识、数据使用和同行应用描述不够的缺陷．

(3) 易于管理．有强大的后台管理功能，容易发布和维护数据．

(4) 着眼未来．强调平台技术的先进性．强调数据中心之间的互操作、数据和知识的存储以及数据挖掘．

(5) 突出服务．提供在线和离线服务，逐渐加强在线服务．系统设计、实现和运行都围绕如何提供更好的服务来开展．

依据以上原则，西部数据中心共享平台设计和实现了以下功能：

(1) 检索和导航模块．提供层次级联结构的数据分类系统；提供基于元数据，和基于全文的数据检索；提供基于数据集序列、数据分类、关键词、时间和空间等数据导航数据．

(2) 元数据模块．在国际标准 ISO 19115 元数据标准的基础上，扩展环境和生态科学的元数据．同时通过数据说明文档增强对数据背景、数据使用和已有同行应用情况等内容的描述．

(3) 数据—知识关联．数据中心除了提供数据和元数据，还包括数据的相关文档、涉及数据的有关文献、使用了这个数据的论文等相关信息．实现了数据和知识的双向链接．

(4) 后台管理模块．多种方式的数据上传和元数据发布(如通过 Web、ArcCatalog、Harvest 等)；

强大的后台管理；代码的版本控制管理、多层体系结构的应用.

(5) Web Service 和数据中心间的集成. 数据中心元数据兼容于 ISO 19115 标准，兼容于 Arc IMS 9.x，提供 WMS(Web 制图服务)等 OGC 建议的服务.

西部数据中心数据共享的特点在于：1)面向科学问题组织数据. 除提供常规数据，西部数据中心还按常用水文、生态模型组织模型数据集、发布特定于研究目的的数据集光盘(比如，中国土地利用数据集光盘，归纳了国际上和国内的土地利用数据集)；2)在线共享与离线服务并重，推送式服务，坚持“全面和开放”(Full & Open)数据政策；3)技术先进的数据共享平台；4)数据与知识关联，数据—文献文档高度集成；5)注重数据的深层应用，重视形成高质量的特色数据集和模型数据集.

1.3 技术方案

在线数据共享平台采用瘦客户端、胖服务器端的浏览器/服务器(B/S)结构. 客户端包括多数流行 Web 浏览器(如 Internet Explorer、Firefox 等)、ArcCatalog(发布和修改需要授权)和 Arc Explorer 客户端.

如图 2，服务器端底层是微软的 SQL Server 2000 Service Pack 4 数据库软件. 西部数据中心维持两个数据库，元数据数据库和主控数据库. 元数据数据库由 ArcIMS 元数据服务通过 ArcSDE 9.0 for SQL Server 2000 创建，并通过 ArcSDE 应用程序接口(API)维护. ArcIMS 集成了 TomCAT 5 Web 服务器用于访问和管理目的. 西部数据中心限制了外部对 TomCAT 的访问，只有授权用户通过 TomCAT 进行元数据存取和管理. 元数据数据库和主控数据库通过 SQL server 的存储过程函数根据元数据文档的唯一指示符(UUID)进行同步.

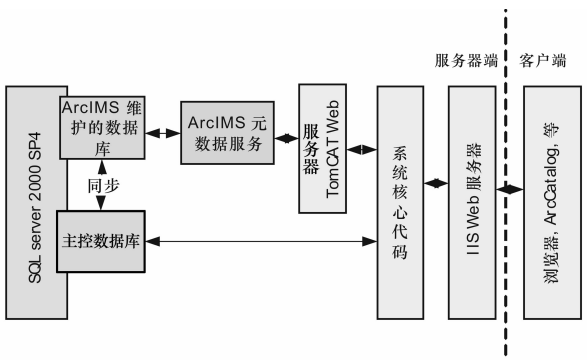

图 2 在线数据共享平台技术方案

Fig. 2 Implementation diagram of the online data sharing sub-system of WestDC

系统核心代码负责 ArcIMS 元数据服务和数据中心其余部分的协同工作.

1.4 多层架构开发模式

西部数据中心在线共享平台在软件开发模式上采用多层架构(n-tiers)，分解为数据库(Database)、数据访问层(DAL)、业务逻辑层(BLL)和表示层(图 3). 数据库层负责具体数据的存储. 数据访问层建立一系列函数，隐藏了具体的 SQL 查询语句. 业务逻辑层特定于具体的业务目的，比如新建元数据，组合一系列的数据访问层函数，形成更高级别的 API. 表示层控制了处理结果的 Web 显示，并控制用户对资源的访问权限.

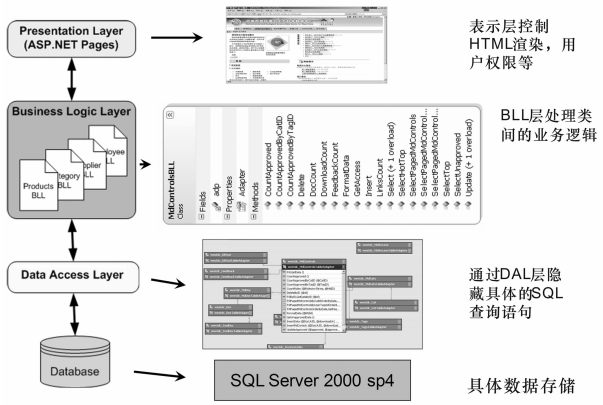

图 3 多层架构的软件开发模式

Fig. 3 N-tier software design pattern employed for WestDC development

多层架构的开发模式有明显的优势，表现在：—1)隐藏了容易出错的 SQL 语句，完全从面向对象角度考虑问题；2)逻辑关系明确，容易定位代码错误；3)高度灵活的代码重用. 然而多层的开发模式也带来了代码冗余的代价. 考虑 Web 应用部署在高性能的服务器上，冗余带来的性能开销是可以接受的.

1.5 数据库设计

西部数据中心数据库包括 ArcIMS 维护的数据库和主控数据库(图 2). ArcIMS 维护的数据库由 ArcSDE 生成，ArcIMS 元数据服务在此库中生成五个新表，对存储的 XML 进行存储和索引.

主控数据库中的数据表按功能分为 7 个包(图 4)，19 个数据表. 核心是元数据(Metadata)包，包括了元数据相关的全部表，负责元数据的管理，并保持与 ArcIMS 数据库的同步. 主控数据库的元数据通过唯一的元数据文档标识符 UUID 与 ArcIMS 数据库连接，建立 vw_westdc_ImportArcIMS 和 wv_westdc_ SyncArcIMS 两个视图表(图 4)，前者显示 ArcIMS 数据库有而主控表没有的记录，后者

显示 ArcIMS 数据库没有而主控表有的记录，此两视图与相关存储过程协同完成 ArcIMS 数据库和主控数据库的同步.

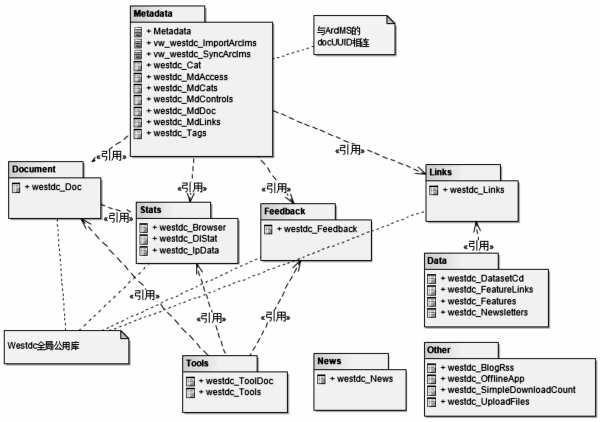

图 4　西部数据中心的数据库 UML 视图

Fig. 4　UML view of the databases implemented in WestDC

文档(Document)包概括了数据文档有关的表；工具(Tools)包概括了数据工具有关的表；新闻(News)包概括了数据共享平台新闻；统计(Stats)包存储元数据、文档、工具等有关统计信息；反馈信息(Feedback)包为元数据、工具提供用户反馈支持；特色数据(Data)包为特色数据集(西部数据中心根据某一主题比如土地利用，将多个数据打包形成一个特色数据集)、西部数据中心简讯提供数据表支持；链接(Links)包为元数据、特色数据集、简讯存储外部链接；其它(Other)包概括了西部数据中心共享平台其它部分的数据表，比如上传的文件信息、离线数据服务的信息等.

2　关键实现

2.1　元数据标准

西部数据中心包括大量的遥感影像、地图等空间数据，在元数据标准的选择上采用 ISO 19115 地理信息元数据标准[16-17]. 考虑到西部数据中心还包括环境和生态学科里的野外观测数据，ISO 19115 不能很好描述这类数据. 我们在 ISO 19115 的基础上进行了扩展. 扩展采用 ISO 19115 文档推荐的九阶段扩展方法[16]进行. 扩展内容主要来自生态元数据国标对野外数据描述的元数据项，比如观测地点、观测内容、数据采集仪器及采集方法的描述等.

元数据标准采用 ESRI Profile 实现. 鉴于 ISO 19139[18]推荐的 ISO 19115 的实现方法，与 ESRI Profile 在命名上不一致，因此，未来西部数据中心元数据标准通过元数据映射(Metadata Crosswalk)的方法移植到 ISO 19139.

出于数据有效性检验、数据中心集成的需要确定的西部数据中心最小元数据项包括：数据标题，数据引用日期，描述数据的语言，数据类别，摘要，元数据联系人，元数据创建日期，描述元数据的语言，数据的地理位置.

元数据以 XML 文档存储在 SQL Server 的 XML 字段里，元数据检索由 ArcIMS 相关的 ArcXML API 实现. 通过 XML 到 HTML 的转换模板(XSLT)实现元数据的显示(图 5).

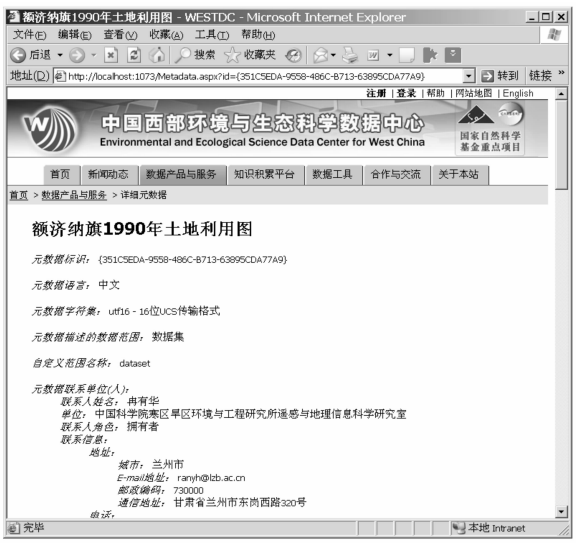

图 5　元数据的页面显示

Fig. 5　An example Web page for displaying metadata using XSLT to convert XML to HTML

2.2　数据-知识关联

越来越多的数据中心实践表明，数据中心不仅存储数据，数据中心还应当包括使用数据所需要的有关知识[19]，这些知识对于有效使用数据是至关重要的，包括数据的背景知识、处理数据有关的工具，以及同行使用这些数据做了哪些研究工作等. 通常的数据中心的做法是对使用数据所需的知识进行一些简单的描述. 西部数据中心在数据和知识关联方面做了良好的尝试，如图 6 所示. 每一个数据有对应的详尽的元数据描述. 每个元数据文档关联一个或多个数据说明文档、辅助使用数据的有关工具及工具文档、以及使用这些数据产出的相关文献. 数据和文献的连接基于主题和关键词的模糊搜索. 文献库来自项目合作方，即中国科学院资源环境信息中心庞大的数字文献资源库.

2.3　分类、检索和导航

西部数据中心采用级联动态分类. 系统类别由

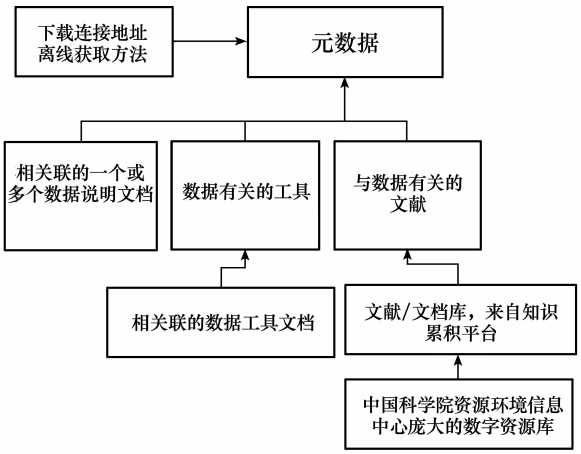

图 6 数据-知识的关联

Fig. 6 Linkage between data and knowledge

管理员从后台添加(图 7). 数据与类别之间是多对多关系，数据可以归类于多个类别. 如果一个数据属于子类别，在父类别里也可见此数据. 这种灵活的分类方式允许从不同的侧面(比如，按数据类别、按科学问题、按主题等)对一数据进行归类，也为后续的多数据平台共享一个元数据库奠定了基础. 事实上，“数字黑河”(http://heihe.westgis.ac.cn/)黑河数据共享平台元数据库是西部数据中心元数据库的子集，其元数据归类于图 7 中的“按主题分类—流域专题—黑河流域”.

图 7 西部数据中心已有分类系统

Fig. 7 A screenshot of the page maintaining the data categories in the WestDC

同时西部数据中心支持用户定义标签. 注册用户可以为每项自己感兴趣的数据自定义标签，以后再登录时便可以快速访问标签指定的这些类别.

西部数据中心提供基于元数据文档的关键词检索和全文检索. 同时提供多种数据导航方式. 当用户面临解决一个科学问题而不知道有哪些数据可被使用时，数据导航可以帮助用户逐个浏览相关的数据. 目前西部数据中心支持基于数据集序列、数据分类、关键词、时间、空间、组合导航以及以上导航方式的组合的导航.

3 结论与展望

“中国西部环境与生态科学数据中心”的网络共享平台，围绕方便查询、详细数据说明、易于管理、着眼未来和突出服务，尽最大努力实现数据共享的“完全与开放”选择功能设计和技术方案.

西部数据中心系统将重视与国内主要数据中心的连接和共享，包括建立元数据收割及互操作机制，互以补充. Web 服务是其中有效的多数据中心集成的方法，比如通过 CSW(目录服务)进行元数据条目的共享和收割.

西部数据中心后续工作将在实现这些标准服务的基础上，进一步发展模型的 Web 服务，提供一系列的模型服务. 模型使用者可以充分利用服务器资源和数据资源，从而将更多的精力集中到科学问题上.

西部数据中心未来的工作也体现在加强数据挖掘的研究和开发，加强数据科学平台和知识积累平台的发展，推进国内数据科学的研究深度，使西部数据中心更充分服务于中国西部环境和生态问题的研究.

参考文献(References):

[1] Klemens B. Modeling with Data: Tools and Techniques for Scientific Computing[M]. Princeton University Press, 2008: 1-470.

[2] Prime Minister's Science Engineering and Innovation Council (pmseic) Working Group. From Data to Wisdom: Pathways to Successful Data Management for Australian Science[R]. Australia, 2006. 1-94.

[3] Committee on Scientific Accomplishments of Earth Observations from Space National Research Council. Earth Observations from Space: The Frist 50 Years of Scientific Achievements[M]. New York: National Academies Press, 2008: 1-144.

[4] Leng Suying, Li Xiubing, Cheng Guodong, *et al*. The progress of studies on the environmental change and ecological issues in western China[J]. Bulletin of National Natural Science Foundation of China, 2005(05): 262-267. [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.]

[5] 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]. Ad-

vances in Earth Science, 2008, **23**(6): 628－637. [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.]

[6] Crompvoets J, Bregt A, Rajabifard A, *et al*. Assessing the worldwide developments of national spatial data clearinghouses[J]. International Journal of Geographical Information Science, 2004, **18**(7): 665－689.

[7] Rishbeth H. History and evolution of the World-Data-Center system[J]. Journal of Geomagnetism and Geoelectricity, 1991, **43**: 921－929.

[8] Meyer T, Suresh R, Ilg D, *et al*. Mosaic, HDF and EOS-DIS: providing access to earth sciences data[J]. Computer Networks and ISDN Systems, 1999, **28**(1－2): 221－229.

[9] Goodchild M F, Fu P, Rich P. Sharing geographic information: An assessment of the geospatial one-stop[J]. Annals of the Association of American Geographers, 2007, **97**(2): 250－266.

[10] Mcpherson R A, Fiebrich C A, Crawford K C, *et al*. Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet[J]. Journal of Atmospheric and Oceanic Technology, 2007, **24**(3): 301－321.

[11] The Investigation Group of the Scientific Data Sharing Project. The overall structure of the Scientific Data Sharing Project[J]. China Basic Science, 2003(01): 63－68. [科学数据共享调研组. 科学数据共享工程的总体框架[J]. 中国基础科学，2003(1)：63－68.]

[12] Qin Dahe. Meteorology data sharing is essential to effective utilization of national resources[J]. China Territory Today, 2003(2): 27－29. [秦大河. 气象科学数据共享是国家资源有效利用的必然选择[J]. 今日国土，2003(2)：27－29.]

[13] Gui Wenzhuang. The 20-year development of CAS Scientific Database [J]. Bulletin of the Chinese Academy of Sciences, 2007(1): 87－89, 91. [桂文庄. 迎接科学数据库发展的新阶段——中国科学院科学数据库发展 20 年的回顾与思考[J]. 中国科学院院刊，2007(1)：87－89，91.]

[14] Webster F. Threat to full and open access to data[J]. Science International, 1997, **65**: 11－12.

[15] Day M. Responsibility for digital archiving and long term access to digital data[J]. Program-Electronic Library and Information Systems, 1999, **33**(4): 383－384.

[16] ISO. ISO 19115:2003[S]. Tc211, 2003:1－140.

[17] ISO. ISO 19115:2003/Cor 1:2006[S]. Tc211, 2006: 1－33.

[18] ISO. ISO/TS 19139: 2007: Metadata-XML schema implementation[S]. Tc/sc, 2007: 1－111.

[19] Seeds F T. Strategic Evolution of Earth Science Enterprise Data System (SEEDS) Formulation Team Final Recommendations Report[R]. USA: NASA, 2003: 1－80.

Design and Implementation of the Online Data Sharing Portal of Environmental and Ecological Science Data Center for West China

NAN Zhuo-tong[1], LI Xin[1], WANG Liang-xu[1], DING Yong-jian[1], ZHU Zhong-min[2], WU Li-zong[1]

(1.Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou Gansu 730000, China; 2.Scientific Information Centre of Resources and Environment, Chinese Academy of Sciences, Lanzhou Gansu 730000, China)

Abstract: The Environmental and Ecological Science Data Center for West China, sponsored by the National Natural Science Foundation of China, is a data sharing portal serving environmental and ecological studies in West China. The overall architecture of this data center will be firstly presented, followed by an introduction of its technical implementation schema, development pattern and underlying database design. Key functional components including adopted metadata standard, knowledge data linkage, and data cataloguing and navigation are examined in detail. This data center is technically designed, leveraging advanced Web and GIS technologies, to meet five principles which are refined upon reviewing existing data centers and considered being important for a successful data center, i. e., easy to find, easy to use, easy to manage, planning for future and highlighting services. Full and open data sharing is technically guaranteed wherever data policy is permitted to share. Finally, the integration among data centers and the issues of modeling services are discussed.

Key words: data center; online data sharing portal; West China; metadata; data science