走进麻省理工大学的DSpace——数字空间图书馆

□ 刘燕权 / 美国南康涅狄格州立大学 康州纽黑文市 美国 06515
高博 / 哈尔滨工业大学国家示范性软件学院 哈尔滨 150001
罗艺 / 中央财经大学信息学院 北京 100081

摘要：DSpace（Digital Space，数字空间）系统由美国麻省理工学院图书馆（MIT Library）和美国惠普公司实验室（Hewlett Packard Labs）共同组建，收录了麻省理工学院教学科研人员和研究生提交的会议论文、学位论文、研究与学术报告、工作论文等，可按院系、题名、作者等方式进行浏览，也可以对任意字段、作者、题名、关键词、文摘、标识符等进行检索，并可在线看到全文。文章对美国麻省理工学院数字空间图书馆（MIT DSpace Library）的成立背景、具体建设及现状进行了综合性的评述，包括项目背景、资源组织、服务特点和技术特征，同时给出了作者的评价与建议。

关键词：数字空间，数字图书馆，麻省理工学院图书馆，惠普公司实验室，开源软件
DOI：10.3772/j.issn.1673—2286.2011.07.011

1 项目背景与项目概述

1.1 应对信息多元化需求与信息供应链改变

随着科学技术的新发展，信息的增长表现在几个方面：信息数量呈指数级的增长，其中数字资源每年以50%－100%的速度增长；信息载体不断增多，尤其以网络为载体的信息；信息格式层出不穷，大体上可以分为文本型、数字型、图片型、多媒体等。随着越来越多的信息资源的出现，如何创建一个稳定的环境来长期保存和获得这些信息成为了一个重要问题。

传统的信息供应链模式是：作者→出版机构→图书馆→读者。随着计算机技术，特别是网络通讯技术的发展，传统的信息供应链模式已经完全被打破。作者可以把稿件发给出版机构，由出版机构将其出版发行；另一方面，作者也可以把稿件或相关资料发到网络上，在网络上进行传播。读者可以绕过图书馆从网络上通过信息服务门户或搜索引擎工具找到所需要的信息。信息供应与信息服务呈现多元化，信息获取也呈现多途径的方式。

传统图书馆职能也因此而改变，现在与未来数字图书馆的功能之一，是要对以分布式网络和计算环境为技术支持和应用基础，对海量的不同类型的信息资源进行制作、组织、存储、发布和检索，以实现读者之间、读者和图书馆专业人员之间的多维和实时沟通。在新的运行环境中，信息活动将从原来的单向转入多向，读者可以自己实现信息的提交、交换和利用。数字图书馆为此将推行新的服务机制和模式。大学数字图书馆可为校园提供整体服务，成为校园资源中心。

新的信息供应链使新的信息活动中的利益博弈产生新的变化。以学术期刊定购为例，各图书馆目前面对两个事实：

一方面，由于学术期刊的出版订费不断上涨，大学图书馆不得不删订期刊的种数或减少图书的购置。根据美国研究图书馆学会（ARL，Association of Research Libraries）的调查，从1986年到2002

年，美国的研究图书馆期刊经费增加了227%，但是订购的期刊种数却只增加9%，而购买图书的种数则下降5%。图书馆供研究使用的资源相对减少，其结果连带影响研究及教学的质量。

另一方面，学者辛勤努力的研究成果发表时，出版社往往向学者按页索费（page charge），当学者需要参考利用期刊时，图书馆却需要再付经费订阅。图书馆需要发展和支持通过开放获取（Open Access）模型发表作品的机制，并向学者、社会提供这种机制。

1.2 开源信息系统开发

针对信息用户在获取信息时总希望能得到“一站式服务”的新特点，即用户希望通过一个检索平台或门户，以最少的操作得到所需要的全部有用信息，不同的研究机构、组织提出了相应的解决方案并开发了不同的应用系统。例如：Eprints（由英国南安普敦大学University of Southampton开发）；Fedora（由Univ. of Virginia and Cornell开发）；GreenStone（由Univ. of Waikato, Witten & Bainbridge 开发）。其中， MIT和HP联合开发的DSpace系统以其高度的灵活性、可用性、可自定义性和开放源代码，受到高校（包括英国剑桥大学、加拿大多伦多大学和美国康奈尔大学、哥伦比亚大学等）和研究机构的青睐和推崇。

Dspace@MIT是一个数字化的知识库，用于保存来自MIT各种形式的学术资源。在创建Dspace之前，MIT的图书馆没有一个保存数字资源的系统。许多教职员工的数字化研究成果都分散在个人网站、个人的电脑和各系的服务器上，这就出现了关于保存和安全的问题。2003年3月休利特-帕卡德公司出资180万美元和MIT合作来开发数字化信息系统，并于2002年9月发布Dspace。

2002年11月4日，MDL采用该系统提供实时服务，为该校学者、教学科研人员和研究生提供了MIT教学科研人员和研究生提交的论文、会议论文、预印本、学位论文、研究与学术报告、工作论文和演示稿全文等宝贵资源。各类数字资料的长期有效管理是资源整合的基础性问题，文档管理系统虽然早已存在，但多数为商业化的组织专有，高校和科研机构往往不得不依赖于供应商。DSpace则提供了运行机构数字资源库所需的基本功能，并且为数据的长期保存及访问提供了发展基础。

MDL吸收了相关领域很多早期的研究和开发成果，比如吸收了Kahn和Wilensky的分布式数字对象服务框架，以及Armsetal在数字图书馆体系结构方面所取得的成果。MDL吸收的另一个重要成果是美国空间数据系统咨询委员会提出的开放档案信息系统（OAIS）参考模型。MDL与南安普顿大学开发的EPrints系统有许多类似的特性，其用户界面吸取了EPrints用户界面设计和使用所积累的经验，特别是借鉴了EPrints的互操作性。MDL目前可以通过元数据采掘OAI协议（OAI-PMH）部分实现互操作，以提供跨资源库的访问服务，方便用户检索资料。

Dspace代表了MIT的一种开放的和学术的文化，它的使命是能够迅速和广泛地分享它的研究和教育成果。目前已有936所院校在Dspace上注册。

Dspace@MIT收集数字化研究报告、数据、图像、音频/视觉影像，数据库和其他来自MIT教职员工和研究人员的其他形式的数字化学术成果。 它也收集教育资料、教师创建的学习资料，如课堂讲稿、考试样题、课程日历、复杂的模拟和可视化、多媒体演示以及演讲录影带。档案包含超过1200余篇学术方面的文章和论文、会议论文25000余篇。MIT估计到2009年最终用户的下载将达到740万篇，平均每天20000篇的下载量。见图1 DSpace@MIT主页。

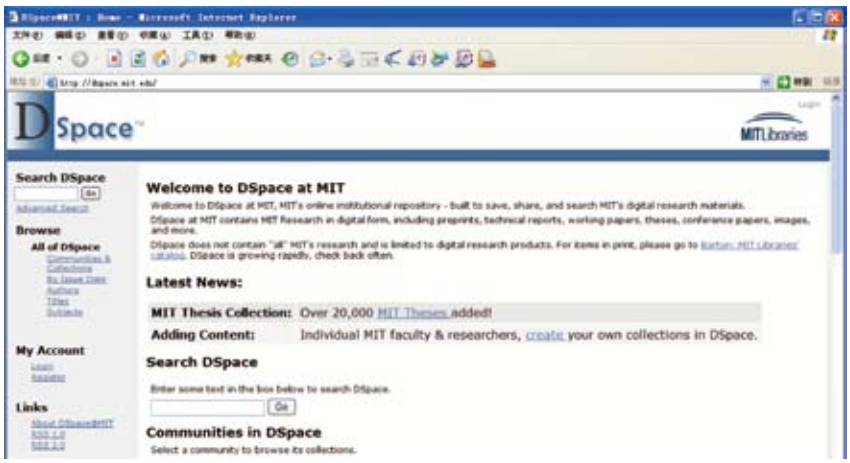

图1 DSpace@MIT主页

2 资源组织与技术特征

MDL作为一个集中的MIT数字图书资源服务来进行运营。MIT中不同的群（community），如MIT实验室、MIT研究中心、MIT研究院系在系统内拥有独立的区域， 见图2。MIT教学科研人员和研究生能够直接通过Web用户界面存放内容，系统尽可能地简化了存放过程。同时系统还提供了批量条目导入程序（batch item importer）作为另一个选择，用于大批量地载入内容。每个MIT教学科研人员和研究生都可以指定专门的人员作为“看门人”，他们将在提交内容被收入到主资料档案库前对其进行审查和编辑，随后MDL系统将索引与数字条目一起提交上来的元数据，根据由MIT教学科研人员和研究生确定的访问权限使其能够被访问。

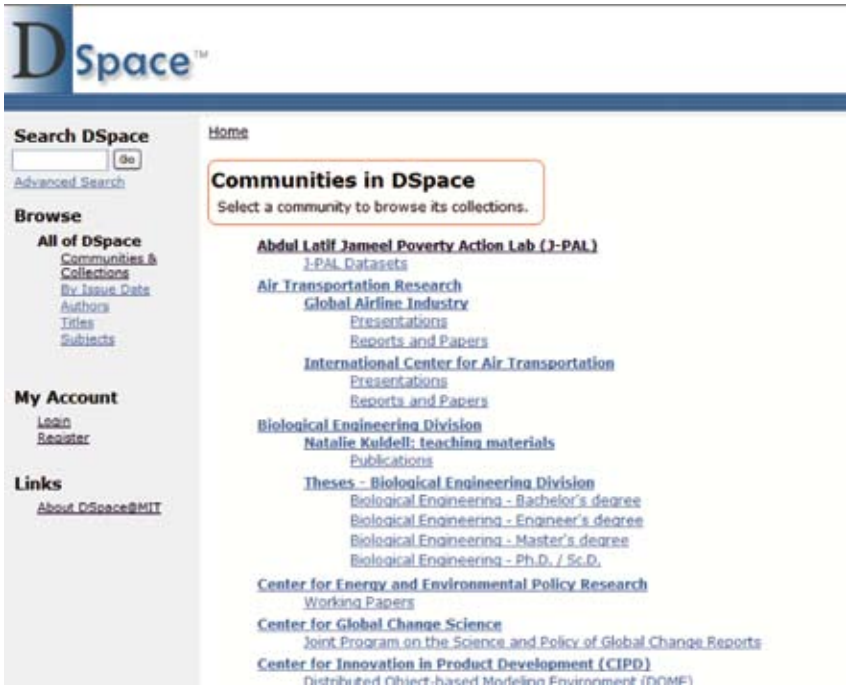

图2 MDL 服务群

2.1 数据模型

DSpace系统的组织结构基于MDL对馆藏数据的组织方式。每个DSpace图书馆站点都被划分为群（community），这些群可以代表MIT的实验室、研究中心或系；群中包含了集（collection），它们是一个相关内容组合；每个集又由条目（item）组成，它们是档案库中的基本存档单元；条目可以进一步被细分为比特流包，通常是普通的计算机文件；关联比较紧密的比特流被组织成包，如组成一个HTML文档的HTML文件和图片，表1是DSpace图书馆中的类型对象及实例举例。

数据模型中一个条目可以属于一个或多个集，一个集也可以位于超过一个的群中，每个条目都拥有限定的元数据记录，其他的元数据可以作为连续的比特流储存在条目中。同时每一个比特流都与一种比特流格式相关联，因为保存服务是DSpace图书馆服务的一个重要方面，所以捕获用户提交的文件的具体格式非常重要。在DSpace图书馆中，比特流格式是指代特定文件格式的唯一并且一致的方式，比特

表1 MDL类型对象及实例举例

对象类型	实例举例	备注
群（community）	MIT计算机科学实验室；MIT海洋学研究中心	
集（collection）	MIT LCS技术报告；MIT ORC统计数据集	
条目（item）	MIT技术报告；MIT讲座的视频记录等	单个记录
包（bundle）	组成HTML文档的一组HTML和图像比特流	
比特流（bitstream）	单个的HTML文件；单个图像文件；一个源代码文件	比特流格式，如Microsoft Word 6.0 版；JPEG编码的图像格式

流格式中不可缺少的一个部分是对该种格式的资料如何进行编译的含蓄或者明确的定义。比如编译采用静态图像压缩的JPEG标准进行编码的比特流，在标准ISO/IEC 10918-1中有非常明确的定义；而编译Microsoft Word 2000格式的比特流，只通过参考Microsoft Word 2000应用进行了含蓄的定义。每个比特流格式都附带有一个支持等级，表示DSpace图书馆在未来以何种格式保存内容。DSpace图书馆为比特流格式分配确定了三种可能的支持等级：

（1）支持的格式，如果格式能够被识别，DSpace就采用各种合适的技术组合，使这种格式的比特流用于数据库中。

（2）已知的格式，如果格式已存在，DSpace既保证比特流保持原状，并允许将其恢复；同时尝试获得足够的信息来将该格式升级到“支持”的等级。

（3）不支持的格式，如果格式不能被识别，DSpace会使比特流保持原状，并允许将其恢复。

2.2 元数据

DSpace图书馆的存档内容有三种类型的元数据：说明性、管理性和结构性元数据。

（1）说明性元数据：每个条目都有一个限定的元数据记录。DSpace图书馆以及与开放源代码共同交付的缺省配置采用的是DSpace图书馆应用文档元素和限定词（qualifier）集的派生，有其他要求的DSpace图书馆可以轻松地对其进行修改，因为DSpace图书馆保留了元素和限定词的注册（registry）。

（2）管理性元数据：包括保存元数据、来源和授权政策数据。该类元数据大部分都保存在DSpace图书馆的关联数据库管理系统（DBMS）语句集（schema）中。

（3）结构性元数据：包括关于如何为最终用户显示条目，或条目中比特流的信息，以及条目各组成部分之间的关系。例如，一个包含多个TIFF图片的论文，每个图片都描述论文的一页。结构元数据则说明每个图片是论文一页的事实，以及TIFF图片/页面的顺序。 见图3。

3 服务特征

MDL收录该校教学科研人员和研究生提交的论文、会议论文、预印本、学位论文、研究与技术报告、工作论文和演示稿全文等，随时更新。读者可按院系机构、题名、作者和提交时间进行浏览，也可以对任意字段、作者、题名、关键词、文摘、标识符等进行检索，可在线看到全文，资源丰富且搜索非常方便。

3.1 服务用户

（1）电子用户（e-people）：DSpace图书馆的用户可能是教师、学生，也可能只是一些计算机系统，因此DSpace图书馆把用户称作“E人”。

（2）工作流（workflow）：DSpace图书馆的运作方式、数字图书材料和资源在被接收之前必须经过事先设定的流转审核的步骤。

（3）信息订阅（subscription）：

Home » Abdul Latif Jameel Poverty Action Lab (J-PAL) » J-PAL Datasets

Women as Policy Makers: Evidence from a Randomized Policy Experiment in India

Show full item record

Citable URI: http://hdl.handle.net/1721.1/39126

Title:	Women as Policy Makers: Evidence from a Randomized Policy Experiment in India
Author:	Raghabendra Chattopadhyay; Esther Duflo
Issue Date:	2007-10-05
Abstract:	This data set uses political reservations for women in India to study the impact of women's leadership on policy decisions. Using a dataset councils in West Bengal and Rajasthan, we compare the type of public goods provided in reserved and unreserved village?s councils. Data provided by GP Pradhans, local villagers, and the 1991 Indian Census
URI:	http://hdl.handle.net/1721.1/39126
Other Identifiers:	hdl:1902.1/USBFNOMLAT
Keywords:	gender, decentralization, affirmative action, political economy, gender, decentralization, affirmative action, political economy

Files in this item

Files	Size	Format
study.xml	1.877Mb	text/xml
womenpolicymakers_census_dta.tab	1.936Mb	text/plain
womenpolicymakers_census.dta	1.348Mb	application/x-stata
womenpolicymakers_parta_dta.tab	52.66Kb	text/plain
womenpolicymakers_parta.dta	68.59Kb	application/x-stata

图3 结构性元数据

用户可以向DSpace图书馆发送订阅请求，以便在新材料到来时收到带有内容提要的E-mail提示。见图4数据资源提交说明。

图4 MDL 数据资源提交说明

3.2 服务方式

MDL支持最终用户以多种方式来查找内容，包括通过外部参考书目，如句柄（Handle），检索一个或多个关键词，通过标题、日期和作者索引浏览等。

检索是MDL查找（discovery）的重要组成部分。用户对Web搜索引擎的期望是非常高的，因此MDL的目标是提供尽可能多的搜索特性。MDL的索引和搜索模块具有非常简单操作方法，支持索引新内容、重新生成索引以及在整个库（corpus）、群或集中进行搜索。

MDL查找的另一个重要机制是浏览。通过这一进程，用户可以查看一个具体的索引，比如标题索引，以及与之相关的导航信息，用于搜索感兴趣的条目。可以浏览的索引包括条目标题、条目发布日期和作者。另外浏览的内容还可以只限定于指定的集或群中的条目。见图5MDL搜索界面。

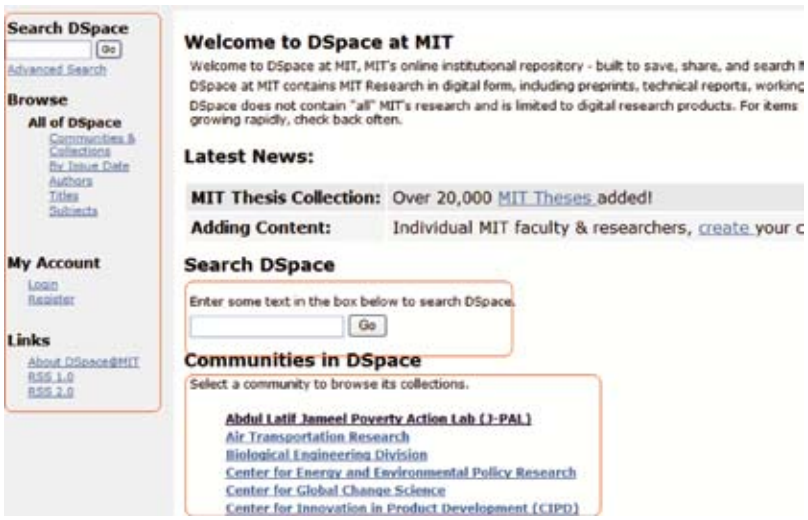

图5 MDL搜索界面

4 技术特性

4.1 开源软件平台

DSpace是一个开放源代码的软件平台，是HP公司和MIT图书馆经过两年的合作开发研制出来的。开发小组与MIT图书馆工作人员以及最初采用该平台的教学人员紧密合作，开发出了“广度优先”（breadth-first）的系统，提供了数字资源库服务所需的各个基本特性。除了作为一种实时服务，DSpace也成为数字资源库功能扩展的平台，特别是用于解决数据长期保存的问题。

Dspace利用最小的额外资源达到学校之间的互通性。这一趋势使得运用主动的开放归档协议来获取元数据。作为OAI的一部分，Dspace的都柏林核心元数据的格式都是可用来兼容的协议。Dspace运行在UNIX平台上，运用其他开源的中间件和Dspace开发人员开发的工具。核心代码是用JAVA语言编写的，其他用到的工具包括PostgreSOL（相关的数据库管理系统），Apache（Web服务器），Tomcat（Servlet引擎），以及Jena（来自HP实验室的工具包）。而DuraSpace.org和社区可能联络有问题，MIT的图书馆和HP公司都不提供给Dspace的用户正式的技术支持。

4.2 检索功能强大

DSpace图书馆的目标就是提供尽量多的检索特色。MDL的索引和检索允许非常方便地索引新内容、重建索引以及在指定范围内检索，最新更新的文章都可以

图6 高级检索界面

在主页上方便地找到。见图6 高级检索界面。

4.3 界面简洁友好

MDL的用户界面都是基于Web的，并且包含一系列的界面：用户提交界面、搜索界面、系统管理界面以及提交审核的各种界面。利用Java虚拟机和Java服务器技术，MDL允许用户通过浏览器访问系统，这使MDL的使用和管理都十分方便。

5 评价和建议——MDL的优点与不足

DSpace系统目前正在全球许多研究机构中应用，作为应对各种不同文献数字化存取的平台工具，如机构知识库（Institutional Repositories，IRs）、学习对象库（Learning Object Repositories，LORs）、电子化主题（eThemes）、电子化记录管理（Electronic Records Management，ERM）、数字化保存（Digital Preservation）、出版（Publishing）等等。

DSpace@MIT最强大的特性之一是简单的最终用户界面。主页清晰整洁，搜索功能很容易发现和使用。MDL具有允许一次登录访问MIT各类分布式信息资源的功能，实现信息资源的可共享性，并可基于自然语言的语义检索对资源进行整合，同时提供学科的个性化服务，基于知识挖掘、过滤、推荐等提供最小化的知识子集等特点。另外，MDL允许数字图书文献在因特网上被免费获取，允许任何用户阅读、下载、复制、传递、打印、搜索和超链接，允许用户在使用时不受财力、法律或技术限制，只需在存取文献时保持完整性，版权归作者所有即可，从而成为众多学者、研究员的资源宝地。

Dspace以一种有组织、易于搜索和分享的方式来收集和归档学术成果。其成功在于其是开源的、免费的、可以更改的、使用方便的、且有利于实现用户需求的组织体系。MIT的长远打算是使Dspace系统在经济运转上有保证和可持续性，这也是它取得成功的重要组成部分。除此之外，全体教员参与到Dspace中也推进了学术期刊改变它的现状，如当前的版权协议不允许在网上发布文章的副本，而这些文献即可在适应它们的地方发表，比如DSpace@MIT归档。而且DSpace@MIT开发者的理想是不仅创造一个MIT的档案，同时也是一个开源的数字图书馆档案软件，可以轻松下载，被全世界的大学所使用。这种资源的共享和学校之间的联合会给未来的学术界带来前所未有的影响。

当然MDL也存在不足之处，通过对DSpace图书馆搭建原理的研究，可以发现当要对MDL系统扩展功能进行比较复杂的修改时，需要修改系统的核心组件，这给图书馆系统的数据库结构和兼容方面带来影响，还有可能导致图书数字资源的丢失或形成乱码。另外是没有可用的在线帮助助理， 而这对于初级和高级的用户来说并不是非常重要的。用户在数字化图书馆的范围内可以利用一定的可用空间来工作和合作是其另一个特色，可以帮助使用者与研究项目更好地结合。其全英文的操作界面并不适合所有人，能提供多国语言版就再好不过了。

参考文献

[1] ATWOOD S. MIT's Superarchive: A digital repository will revolutionize the way research is shared and preserved [J / OL]. Technology Review,2010(1).[2010-09-16]. http://www.technologyreview.com/business/13015/page1/.
[2] BAUDOIN P, BARNSCHOFSKY M. Implementing an Institutional Repository: The DSpace Experience at MIT [J]. Science & Technology Libraries, 2004,24(1/2):31-45.
[3] DSpace@MIT. What is DSpace@MIT? [OL]. [2010-09-16]. http://libraries.mit.edu/dspace-mit/about/definition.html.
[4] DSpace@MIT. Home [OL]. [2010-09-18]. http://dspace.mit.edu/.
[5] DSpace.org. DSpace registry [OL]. [2010-09-16]. http://www.dspace.org/index.php?option=com_formdashboard&orderby=InstNameASC&page=19&Itemid=151.
[6] DURANCEAU E. The "Wealth of Networks" and Institutional Repositories: MIT, DSpace, and the Future of the Scholarly Commons [J]. Library Trends, 2008, 57(2):244-61.
[7] MIT Faculty Open Access to Their Scholarly Articles [J]. Advanced Technology Libraries,2009, 38(5):2.
[8] SMITH M, BARTON M, BASS M, et al. DSpace: an open source dynamic digital repository [J]. D-Lib Magazine,2003,9(1).

作者简介

刘燕权，毕业于美国麦迪逊大学信息学院，现在南康涅克州立大学任终身正教授。近年的出版物主要集中在数字图书馆、数据挖掘、软件工程、信息存储、数字化及多媒体技术、国家信息基础结构、图书馆统计及管理等。E-mail: liuscsu@gmail.com

Dspace @ MIT

Yan Quan Liu / Southern Connecticut State University, New Haven, CT, USA, 06515
Gao Bo / The National Pilot School of Software, Harbin Institute of Technology, Harbin, 150001
Luo Yi / School of Information, Center University of Finance and Economics, Beijing, 100081

Abstract: Dspace@MIT is an institutional digital repository created to collect, organize, preserve and distribute the digitally born intellectual work product of faculty and researchers at the Massachusetts Institute of Technology. As open source software used by the organization with minimal additional resources, DSpace@MIT is also an expression of MIT's culture of openness and the institution's mission to share its research and educational products quickly and widely. This article provides an overview of the project's background , site construction, resources organization, technological features and services provided. Authors' comments and suggestions are also given.

Keywords: Digital library, Dspace, MIT Library, HP Labs, Open Source

（收稿日期：2011-06-09）