

●李春旺

网络环境下学术信息的开放存取

摘要 开放存取思想迎合了网络时代信息交流的特点,开创了一种新的、高效的交流模式。开放存取包含两层含义:一是指学术信息免费向公众开放;二是指学术信息的可获得性。网络环境下的学术信息开放存取服务有多种类型,如:在线文档发布、科学家与研究组内部的直接交流、通过电子邮件的自由辩论、搜索引擎的索引与检索服务、第三方集成检索服务等。图3。参考文献10。

关键词 网络环境 学术交流 学术信息 开放存取 体系结构

分类号 G250

ABSTRACT The conception of open access meets the characteristics of information communication in the network age, and opens a new and efficient communication pattern. Open access has two meanings, i. e. the free access of scholarly information to the public and the availability of scholarly information. 3 figs. 10 refs.

KEY WORDS Networked environment. Scholarly communication. Scholarly information. Open access. Infrastructure.

CLASS NUMBER G250

1 网络化数字化环境下学术信息的开放存取

开放存取先导对开放存取是这样定义的:所谓开放存取(Open Access, OA),是指通过公共网络可以免费获取所需要的文献,允许任何用户读取、下载、拷贝、分发、打印、检索以及获取在线全文信息,支持爬行器收获并建立本地索引,支持用于法律允许的其他目的(不包括商业、法律、技术贸易壁垒方面的应用),惟一的限制是复制与发行^[1]。开放存取包括两层含义:一是指学术信息免费向公众开放,它打破了价格障碍;二是指学术信息的可获得性,它打破了使用权限障碍。开放存取服务包括多种类型,如在线文档发布、科学家与研究组内部的直接交流、通过电子邮件的自由辩论、讨论组及相关服务、搜索引擎的索引与检索服务、第三方集成检索服务等。促成开放存取机制的建立有多方面的因素。

(1)基于传统出版方式的学术交流模式面临越来越多的挑战。①科研成果迅速增长使有限的印刷出版能力变得愈加不适应,用户无法忍受印刷出版过长的滞后性。②印刷型期刊及其电子版本的价格不断攀升,加重了图书馆经费预算的困难。③由谁负责保存数字化学术成果资料的问题变得越来越不确定。④基于印刷型文献的学术交流具有内容单一、出版周期长、流通渠道不畅、可获得性差(如付费使用)等缺点,影响了信息的正常交流。

(2)网络环境下科学研究呼唤新的学术交流机制。①研究机构需要保存自己产生的知识信息及知识产权,包括尚未在期刊上发表的知识。②科研人员希望能全面、准确、及时地获取最新的科研进展、成果与思想。③科研机构希望扩大自己研究成果的影响力。

(3)网络环境下科研人员信息交流存在很多障碍。如网络平台、媒体类型、信息格式处理与转换等技术障碍,学术信息搜索、获取、服务等管理障碍。

针对以上问题,人们提出了开放存取机制。它通过建立机构数据仓库系统,借助先进的信息组织与知识管理技术,收集机构内部产生的各种学术信息,以网络方式出版、发布,建立开放存取联盟,辅助功能强大的搜索引擎揭示服务、联邦检索服务等,最大限度地推动科研成果信息的传播与交流,使科研人员可以及时、方便、快捷地获取所需要的信息。

2 开放存取——网络时代学术信息交流的新模式

开放存取思想迎合了网络时代信息交流的特点,开创了一种新的、高效的交流模式。目前对开放存取的研究,呈现出蓬勃发展的势头。

(1)布达佩斯开放存取先导(Budapest Open Access Initiative, BOAI)。为解决学术成果出版、发布问题,促使学术信息在因特网上广泛自由使用,OSI(Open Society Institute)于2001年12月布达佩斯会议

上提出了开放存取的思想并创立了 BOAI。BOAI 包括两个主要目标:一是建立学术信息自管理系统 (Self-Archiving)^[2],即通过建设研究机构数据仓库 (Institutional Repositories),搜集、组织、保存、发布机构内生产出的知识信息,并向全社会提供免费的共享服务。二是建立开放存取期刊 (Open-Access Journals, OAJ),BOAI 通过要求作者放弃稿费的方式,保证期刊免费向社会开放;通过要求作者拥有著作权或将其转让给出版者的策略,保证著作在网络环境下开放存取的合法性;通过建立专家评审制度,保证开放存取期刊论文的质量。

(2) 相关研究项目。开放存取运动吸引了众多组织机构,如美国国会图书馆, Harvard Virginia Tech, Los Alamos, Cornell 大学, CNI (Coalition for Networked Information), NSF (National Science Foundation), Mellon 基金会等,著名的研究项目有 SHERPA, SciX, Dspace, Eprints 等。① SHERPA (Securing a Hybrid Environment for Research Preservation and Access) 从技术、管理以及文化角度,研究网络出版数据库的建设问题,包括知识产权保护 (IPR)、文献质量控制、开放存取、互操作以及网络出版资源的保护等^[3]。② SciX 利用商业过程分析技术,调查当代商业出版模式特点,提出适合科学交流需求的新的商业模式。③ Eprints 是英国 Southampton 大学设计的、面向研究机构学术信息的开放存取系统,它兼容 OAI 技术,使学术信息更容易被获取,目前已经有 132 家用户。④ Dspace 是美国麻省理工学院 (MIT) 和 HP 公司联合开发的一个开放源码系统,它实现了对大学科研智力成果信息的获取、发布、永久性保存以及开放访问服务。它采用数字对象技术,实现对文本、音频、视频、图像等多种信息的存储与内容管理:采用 OCLC 的 OAICAT 技术,实现对 OAI-MHPv2 的兼容;采用 CNRI 的 Handle 系统来建立惟一标识符框架。⑤ 其他项目还有 Fedora、CARL 等^[4]。

(3) 开放存取期刊。开放存取期刊发展的一个显著特征是:传统出版商纷纷加入到开放存取期刊行列。2003 年 5 月 12 日,在 OSI (the Open Society Institute) 的支持下, Lund 大学图书馆与 SPARC (The Scholarly Publishing and Academic Resources Coalition) 联合创建了 DOAJ (Directory of Open Access Journals),其宗旨是增加开放存取学术期刊的透明性、可用性、易用性,提高期刊的使用率,扩大学术成果的影响力,现有开放期刊 1072 种。2003 年 10 月 22 日,德国中

心研究院与其他几个研究院联合发表了“柏林宣言”,宣布将自己的科学发现提供给开放存取工程,并鼓励在科学出版领域推行开放存取机制^[5]。2003 年 11 月 3 日,英国牛津大学出版社宣布与牛津大学图书馆合作参加 SHERPA 项目,允许全球科研人员在线免费搜索访问 2002 年以来牛津大学作者出版的学术论文。斯坦福大学的 High Wire 出版社将其出版的期刊提供全文免费网络服务,截止到 2004 年 5 月 2 日已开放全部 1741023 篇论文的 716615 篇,被称为全球最大的免费全文学术论文数据库。

(4) 当前开放存取研究工作的特点。① 得到了广泛重视,很多国家纷纷建立自己的开放存取研究项目,国际化开放存取组织开始出现,并在数据库的互操作、组织管理、协调指导以及未来发展趋势等方面展开研究。② 内容覆盖众多学科,既有综合性的资源数据库,又有专门的学科数据库;地域上覆盖面广,形成了全球化的发展趋势。③ 形成了一定规模的在线文献,并表现出迅猛的增长势头。④ 网络出版文献质量不断提高,老牌大型检索工具 (如 CA、SCI) 开始收录网络出版文献,并且收录量迅速增加,如 ArXiv 被 SCI 收录的论文已经超过 10 万篇。⑤ 开放存取系统与相关信息系统的集成引发一系列新的研究热点,如与图书馆系统的集成使多种媒体信息类型管理与显示、新的数字对象元数据标准等成为研究热点,与 Portal 的集成则侧重 Web 内容管理的研究,而研究机构数据仓库系统之间的集成以及开放存取联盟的建立,使元数据收获、联邦检索成为研究的焦点。

3 开放存取实现方式

3.1 基于 OAI 的开放元数据机制

OAI (Open Archive Initiative) 对开放存取产生了深远的影响,它最初起源于电子出版界的互操作项目,目的是开发并推广开放互操作协议标准,实现高效的信息交流与传播。因为元数据格式过多,系统间互操作时元数据格式转换和匹配是一个很大的障碍,为此, OAI 指定 DC 作为统一的元数据标准,并以此作为系统的统一元数据接口规范。在开放存取方面, OAI 将资源与服务分离,提出数据提供者、元数据收获与服务提供者的组织机制 (体系结构见图 1)。

数据提供者的主要目标是开放自己的数据,系统遵循 OAI 标准,建有标准元数据访问接口,支持基于 OAI-PMH 的元数据收获。著名的系统有 arXiv、

CogPrints、Virginia Tech 大学学位论文数据库 ETD Individuals 等。

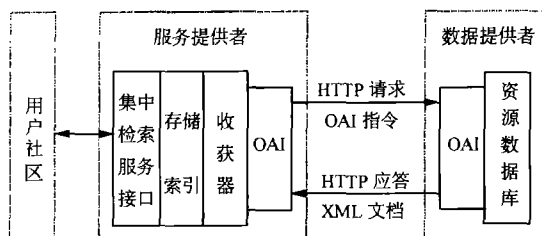


图1 基于 OAI 的开放存取机制

服务提供者提供两种检索服务,即联邦检索(Federate Search)与集中检索。联邦检索采用同步方式,将用户检索请求同时提交给多个 OAI 数据源,各个数据源独自进行检索操作,并返回检索结果。集中检索利用 OAI 收获器,将分散的 OAI 资源元数据收获到本地,并建立本地存储、索引系统,提供集中的检索和全文链接服务。著名的服务提供者有 Southampton 大学的 Eprints、Old Dominion 大学的 ARC、OCLC 的 OAI Cat 和 FS Consulting 的 my.OAI 等。

元数据收获实现数据提供者与服务提供者之间的协议解释与转换。收获器遵循 OAI-PMH 协议,以 POST 或 GET 方式将 OAI 请求提交给数据提供者;数据提供者对请求进行处理后,返回 XML 格式的文档信息;收获器对 XML 文档进行解析后,生成本地存储格式的数据,并建立本地索引,提供集中检索服务。

3.2 基于 DOI 的永久性保存与利用机制

开放存取的一个重要目标是学术资源永久性保存和利用。为实现这一目标,需要为每个资源分配一个永久性的、惟一的标识符,又称数字对象标识符(Digital Object Identifier, DOI)。用户通过 DOI 可以在任何时候,惟一获得指定的信息对象,从而保证用户创建的书签、个性化链接等信息永久有效。DOI 管理系统一般包括 5 部分:命名域(namespace)、惟一标识符、命名机构(Naming Authority)、命名登记机构(Registry)、地址解析系统(Resolution System)^[6]。当前,在采用 DOI 管理框架的开放存取系统中,Dspace 具有一定的代表性。

Dspace 利用 CNRI Handle System 创建全球惟一的数字对象标识符,尽管 CNRI 的 Handle 系统没有在 IETF(Internet Engineering Task Force)注册,但它与 IETF 的 URN(Uniform Resource Name)规范完全兼容。一个

Handle 系统包括 3 项主要功能,即 Handle 的分配、管理与解析。Dspace 采用 Handle 系统为每个信息条目(item)、资源集(Collection)和社区(Community)创建永久性标识符。为实现 Handle 的创建与管理,每个使用 Dspace 网站都必须从 CNRI 获得一个 Handle 前缀,在指定前缀下为自己的资源分配标识符,以保证标识符的惟一性。Handle 标识符有两种表现形式:

hdl:1721.123/4567

http://hdl.handle.net/1721.123/4567

这两个标识符指向同一个资源。第一种方式适合描述物理存放位置永久不变的资源,它需要使用专门的客户端软件(Handle Resolver)对 Handle 进行解析;第二种方式将 Handle 包装成 URL 形式,再通过代理服务器将 URL 映射为本地具体资源。Dspace 采用第二种方式,它更易于 Web 浏览器的兼容,同时可以充分利用 HTTP 协议支持 URL 重定向的特性,当资源在长久保存过程中存放位置发生变化时,通过重定向机制,用户仍然可以找到该 Handle 初始指向的资源。

3.3 基于搜索引擎的开放存取机制

搜索引擎是网络环境下学术交流的重要渠道。一个信息资源是否在搜索引擎的覆盖范围内,将决定它的利用率,进而决定该信息内容的影响力。目前,大部分学术文献还不能达到它应有的影响力,因为它们作为动态资源被存储在数据库中,对搜索引擎是不透明的。基于搜索引擎的开放存取就是要打破传统搜索引擎的限制,让普通搜索引擎能够对数据库内学术资源(Deep Web)进行爬行、下载、索引,并提供广泛的检索服务。Old Dominion 大学 DP9 项目专门研究如何将数据库中的资源贡献给搜索引擎。

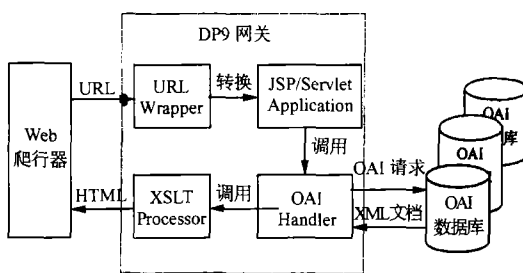


图2 DP9 体系结构

DP9 是一个开放源码的研究项目,它通过建立一个专门的爬行器网关,采用重定向机制,实现了通用搜索引擎(如 Google)对 OAI 数据库资源的爬行、下

载、索引与检索服务(体系结构见图 2)^[7-8]。DP9 包括 3 个组件:URL 包装器(URL Wrapper)、OAI 处理器(OAI Handler)、XSLT 处理器(XSLT Processor)。其工作原理是:①URL 包装器从搜索引擎接受 URL 请求,调用内部的 JSP/Servlet 应用模块对 URL 进行包装,将包装后的请求信息转给 OAI 处理器;②OAI 处理器代表爬行器向 OAI 数据库发出 OAI 请求,并返回 XML 格式的结果信息;③利用 XSLT 处理器,将 XML 格式的结果信息转换为 HTML 格式,响应搜索引擎的请求。DP9 为每个 OAI 数据库定义一个爬行器入口网页(entry page),爬行器通过该网页中的链接就可以爬行到 OAI 数据库中的所有数据。

3.4 基于 Web Service 的开放存取机制

在开放学术信息交流环境下,Web Service 是指由研究机构发布其在线信息服务项目,允许其他机构、合作伙伴以及终端用户通过因特网开放搜索、识别、获取以及动态调配这些在线服务,实现服务层次上的资源共享与互操作。网络服务的核心内容是元数据登记机制,登记规范内容包括系统元数据、UDDI (Universal Description, Discovery, and Integration)、数据库目录(如相关的 DBMS/SQL)、XML 知识库、相关工具、应用部件、Ontology 等。网络服务登记系统应包括 3 项基本功能:网络服务规范描述与登记、存储管理以及服务发现与调用等(体系结构见图 3)^[9]。

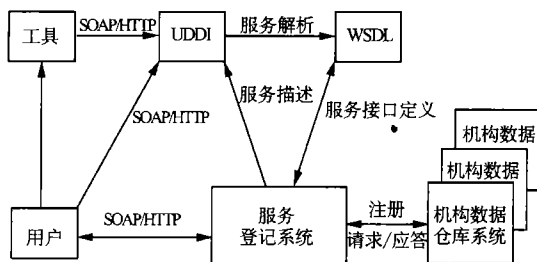


图 3 基于 Web Service 的开放存取模式

(1)对网络服务的描述。采用 WSDL 规范对登记的网络服务进行描述,包括服务提供机构信息(如地址、名称、联系方式)、地点信息(如 URI)、数据交换模型(响应/请求、请求/响应、单向、多向广播等)、输入/输出信息类型(是面向文档的还是面向过程的)、服务所在位置(服务提供端点)、信息格式、错误信息等。为实现网络服务的搜索、定位,要求登记服务的调用接口必须遵循 UDDI 规范,并将服务描述信息在 UDDI 集成与发现系统中注册。

(2)存储管理。被提交到服务登记系统的网络服务一般采取等级式结构存储方式,包括四级结构:服务登记机构、服务提供机构、一组相关网络服务集合以及网络服务。其中,网络服务是四级结构的基本单元,用 UDDI 的 tModel 元素描述。相关的分类标准有 Standard Industry Code、United Union/Standard Product and Services Classification 等^[10]。

(3)网络服务发现与调用。公共用户利用 UDDI 实现对网络服务的检索、发现,获取相关的 URL 信息,并通过 WSDL 的解析,实现对具体机构数据仓库系统提供服务的调用。在知道网络服务内容的情况下,用户也可以直接向服务登记系统发出请求,调用相关服务,获取结果信息。

4 开放存取面临的挑战

开放存取代表着未来信息交流的发展方向,作为新生事物,它也面临众多挑战。

(1)技术问题。包括:①适合海量存储与开放访问的 IT 基础设施建设问题。②如何建立多级的、全球化的联邦信息组织与服务体系结构。③如何构建适合 e-science、Open Grid 发展需要的信息服务模式。④如何突破单一的、静态的信息组织方式,建立复合的、动态的数字对象管理体系,实现知识管理与语义检索服务。⑤技术标准问题。⑥知识产权保护技术问题等。

(2)管理问题。包括:①如何处理作者、出版者、用户和图书馆员等多种角色之间的关系。②如何确立传统期刊向开放存取期刊过渡策略与方法,在不收订购费与访问费的情况下,如何解决网络出版的经费问题。③如何制定网络出版的质量控制策略,特别是如何建立行之有效的、跨机构的评审联盟制度等问题。

(3)知识产权问题。①如何解决传统出版模式向网络出版过渡期间知识产权划分的问题。②网络出版模式下,知识产权的表现形式是什么,如何指定相应的管理策略,如何保护产权人利益等问题。

参考文献

- 1 Budapest Open Access Initiative, FAQ.
<http://www.earlham.edu/~peters/fos/boaiFAQ.htm>. 2004 - 04 - 25.
- 2 Self-Archiving. <http://www.ecs.soton.ac.uk/~harnad/Tp/nature4.htm>. 2004 - 04 - 25.

- 3 Mailing List SPARC-OAForum@arl.org Message 170.
http://mx2.arl.org/Lists/SPARC - OAForum/Message/170.
html. 2004 - 05 - 02.
 - 4 Alan McCord. Institutional Repositories: Enhancing Teaching Learning, and Research.
http://www.educase.edu/ir/library/pdf/DEC0303.pdf.
2004 - 05 - 02.
 - 5 Open Access Conference-Berlin Declaration.
http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.
html. 2004 - 05 - 02.
 - 6 毛军, 张晓林等. URI 和数字对象唯一标识符. 现代图书情报技术, 2003(2)
 - 7 Xiaoming Liu, Kurt Maly, Michael L. Nelson. DP9: An OAI Gateway Service for Web Crawlers. 见: JCDL'02, Oregon, 2002. http://www.cs.odu.edu/~liu_x/dp9/dp9.pdf. 2004 - 04 - 30.
 - 8 DP9: An OAI Gateway Service for Web Crawlers.
http://www.language-archives.org:8082/dp9/about.jsp. 2004 - 05 - 02.
 - 9 Stans Kleijnen and Srikanth Raju. An Open Web Services Architecture.
http://www.acmqueue.com/modules.php?name=Content&pa=printer_friendly&pid=26&page=1. 2005 - 05 - 02.
 - 10 王昉. 数字图书馆的分布式网络服务登记系统. 中国信息导报, 2003(5)
- 李春旺 中国科学院文献情报中心博士研究生, 副教授. 通信地址: 北京北四环西路 33 号. 邮编 100080.
(来稿时间: 2004 - 05 - 20)

(上接第 32 页)

解决这一矛盾的方法是在指导思想以为公众服务提供高水平服务为导向, 在实施的战略上应用现代知识管理原理和技术。

我国各级政府都面临以信息技术为载体、以市场化为动力的全球化浪潮的冲击和以知识创新为内核、以产业信息化为重要特征的知识经济的挑战。在电子政府的建设中, 应该遵循知识管理的思想和原则, 在知识的产生上讲求创新性和高质量, 在知识管理上力求系统积累、方便检索利用, 在知识的使用上追求整个政府系统或全社会共享, 在知识的运用上采用智能化和集约化。电子政府领域内的知识管理可以极大地推进用电子政务服务模式, 改良旧的政府组织结构, 优化政府工作流程; 可以提高政府核心业务能力, 最终提升政府的竞争力, 也提升整个国家社会的竞争力。

从我国电子政府建设和应用情况来看, 我国未来电子政府系统将以知识管理为核心, 可能具有以下功能和特点。

(1) 具备强大的知识信息处理功能。面向知识管理的电子政府自动化系统将集成各种知识信息数据。这些知识信息不仅包括电子邮件信息, 而且还包括文件系统中的文件、传统的关系型数据库数据、数据仓库中的数据, 以及互联网上的各类数据。

(2) 面向社会的具备多种知识共享方式和强大的

知识共享能力。充分利用各种协同工作手段, 包括多线程讨论、文档共享、电子邮件及一些辅助工具, 提供在线及时共享等。将提供不同层面的信息共享方式, 包括移动通讯设备的支持、手机的 WAP 接入访问、PDA 的支持等。电子政府将成为社会知识共享工程的主体部分。

(3) 电子政府内网系统部分可以发挥政府范围内知识管理的平台作用, 逐步把知识管理原则与实践融进每个政府公务员的日常工作, 模拟和实现政府各部门工作流的自动化, 更进一步实现政府多部门交叉工作流中各类相关知识和方法的综合运用。从目前实现的手工作业精简、文档电子化、公文自动流转等, 到重组后台办公流程和改变组织结构原有模式桎梏, 向知识共享、决策辅助等纵深方向发展, 从而最终体现政府业务流程优化、管理和决策水平提升。

参考文献

- 1 姜伟. 电子政务. 知识管理是新焦点. 中国计算机报, 2004 - 02 - 11

刘 辉 浙江财经学院副教授, 博士生. 通信地址: 杭州市. 邮编 310012.

周慧文 浙江财经学院财政与公共管理学院副教授, 武汉大学信息管理学院博士生. 通信地址同上.

(来稿时间: 2004 - 05 - 23)