第25卷 第2期
2006年4月

情 报 学 报

Vol.25, No 2
April, 2006

IR:现状、体系结构与发展趋势

李广建[1] 黄永文[2] 张 丽[1]

(1.北京师范大学管理学院, 北京 100875; 2.中科院文献情报中心, 北京 100080)

摘要 简要介绍了IR的概念和现状,在此基础上对IR的体系框架、功能及主要技术进行详细的论述,最后分析了IR的发展趋势。

关键词 机构仓储 数字对象 开放存取

IR: Current Situation, Architecture and Tendency

Li Guangjian[1], Huang Yongwen[2] and Zhang Li[1]

(1. *Beijing Normal University Management College, Beijing* 100875;
2. *Library of Chinese Academy of Sciences, Beijing* 100080)

Abstract The paper introduces the definition and current situation of Institutional Repositories, and analyzes the architecture, functions and major techniques of Institutional Repositories. At last, it points out the tendencies in Institutional Repositories.

Keywords Institutional Repositories, digital object, open access.

1 IR的现状

Institutional Repositories是近年来出现的一个新概念,目前国内尚没有通用一致的用法,大致可以翻译为"机构仓储"、"机构资源库",为方便起见,本文使用其英文简称"IR"。

Raym Crow于2002年首次将IR定义为:获取和保存一个或多个大学的智力产出的数字化集合。他认为IR主要解决两个问题:一个是改革学术交流系统,强调学术团体对其知识资源的控制权,削弱期刊出版商对学术资源的垄断,降低学术机构和图书馆获取文献的经济成本,方便科研人员获取学术资料;另一个是解决大学质量的评价问题,在当前的学术交流体系下,研究人员学术论文的价值和影响力在一定程度上反映了研究人员及其所在机构的科研水平,IR集中了本单位科研人员的智力作品,能较全面地揭示众多科研成果的科学、社会和经济价值,因而反映了机构的学术质量,同时提高了机构学术成果的透明度和影响力。

随着研究与实践的深入,学术界对IR的认识也不断深入。2003年,网络信息联盟的执行理事Clifford A. Lynch对IR的定义是:一个大学向其成员提供的、用以管理和传播该大学及其成员所创造的电子资料的一系列服务。这个定义显然要比Raym Crow的定义更符合IR的实际。已经不再从"资源"的角度来限定IR,而是将IR看成了一种服务。Clifford A. Lynch对IR的看法被学术界广为接受。

2004年6月24日在伦敦召开了"IR及其对出

收稿日期:2005年4月25日

作者简介:李广建,男,1963年生,北京师范大学管理学院教授,中科院文献情报中心博士生导师,主要研究领域为:信息资源管理、管理信息系统和数字图书馆。黄永文,女,1975年生,中科院文献情报中心在职博士,主要研究领域为:网络信息管理技术与信息系统。张丽,女,1981年生,北京师范大学管理学院在读硕士,主要研究领域为:信息管理技术和信息系统。

版业的影响"研讨会,来自英、美等国的100余位研究人员和实际工作者参加了这次会议,其中包括Raym Crow和Clifford A. Lynch等人。这次会议对IR及其在学术出版领域的作用和影响进行了更为深入的探讨,并对IR给出了较为一致的看法:IR是基于机构的服务,对机构所创建的内容进行存储、传播、管理和监管。

从以上定义中可以看出,IR是在信息化、网络化环境下,为方便学术资源存取、促进学术交流而提出的,是获取、长久保存以及管理来自一个或多个学术团体的知识产品并将其提供给用户访问的一种数字化信息及其服务的集合。

从实践上看,IR既是一种理念,也是一种系统。作为一种理念,IR是对传统学术交流体系的挑战,它提出了一种真正为科学研究服务的、基于网络环境的分离式学术出版模式,将学术出版过程中的采集、加工、出版等环节逻辑分离,使得作者、图书馆员、出版商在学术交流中的角色重新定位。作为一种系统,IR提供了一个方便的平台,使作者能够快捷地提交和发布论文和其他研究成果,同时,使得用户能够方便、快速、无阻碍地获取所需要的学术资源,缩短学术交流周期,提高科研效率。

IR提出以后,受到了学术领域的广泛关注,美国和英法德等欧洲国家都投入巨资进行IR研究,目前全世界已有数百个大学、研究机构和学术团体建立了自己的IR。许多商业公司如HP、Innovative等也提出了自己的解决方案。目前,IR解决方案主要可以分为4类:①专用系统,这类系统是IR研究项目的成果,如eScholarship、JISC IE、Knowledge Bank等;②开放源码和免费的系统,如Dspace、Fedora、Archimede、CDSware等;③商业系统,如Documentum、Bepress、UMI/ProQuest研制的DigitalCommons、DiMeMa公司研制的CONTENTdm、Innovative公司的DRM、BioMed中心的Open Repository等;④混合型的系统,如VTLS公司的Vital等。

为了便于IR建设者选择现有的IR解决方案,美国开放社会研究所(Open Society Institute,OSI)近年来定期发布了IR软件指南,在其2004年8月发布的指南第三版中,列出了Archimede、ARNO、CDSware、DSpace、Eprints、Fedora、i-Tor、MyCoRe和OPUS共9个IR,并从基本情况、技术细节、仓储和系统管理、内容管理、用户接口和查询功能、存档、系统维护等7个方面对这些系统进行了详细的对比研究。据不完全统计,仅这9个系统,全世界就超过了200个用户(机构),见表1。

表1 主要IR的国别分布表(+表示以上)

系统名称	使用范围	使用机构数量
Archimede	加拿大	1
ARNO	荷兰	7
CDSware	欧洲 美国	7+
DSpace	全世界	20+
Eprints	全世界	140
Fedora	全世界	20
i-Tor	荷兰	30
MyCoRe	德国/瑞典	10
OPUS	德国	37

2 IR的体系架构及主要功能

2.1 IR的体系架构

就现有的IR而言,它们一般由三层结构组成,依次为存储层、业务逻辑层和服务层。IR的体系架构如图1所示。

存储层主要是保存数据和对数据进行读、写、删除操作,数据包括数据流和数字对象的元数据包文件。在存储层中,每个数字对象不仅保存有数据流,还封装有元数据包文件。此外,存储层中还使用关系数据库或XML文件来辅助管理数字对象。

业务逻辑层负责执行整个系统的业务逻辑,一般由三个子系统组成,即内容管理子系统、存取子系统和管理子系统。内容管理子系统包括数字对象管理和唯一标识符(PID)生成等模块,前者主要负责数据对象的操作和对象完整性校验,后者负责唯一标识符PID的自动生成工作。存取子系统主要包括数字对象映射和数字对象分发等模块。管理子系统主要包括用户和仓储的安全管理、权限管理、历史日志管理、工作流管理等模块。

服务层主要由Web服务接口、Web UI、OAI-PMH元数据提供服务、联邦服务等模块组成。Web UI模块提供数字资源的提交以及对机构库的浏览和检索。OAI-PMH元数据提供服务模块允许外界按OAI-PMH协议来获取IR中的元数据。联邦服务模块包括一些只为联邦成员提供的服务项目。Web服务接口模块将业务逻辑层以Web服务的形式提供给外界,允许外部系统集成这些Web服务。

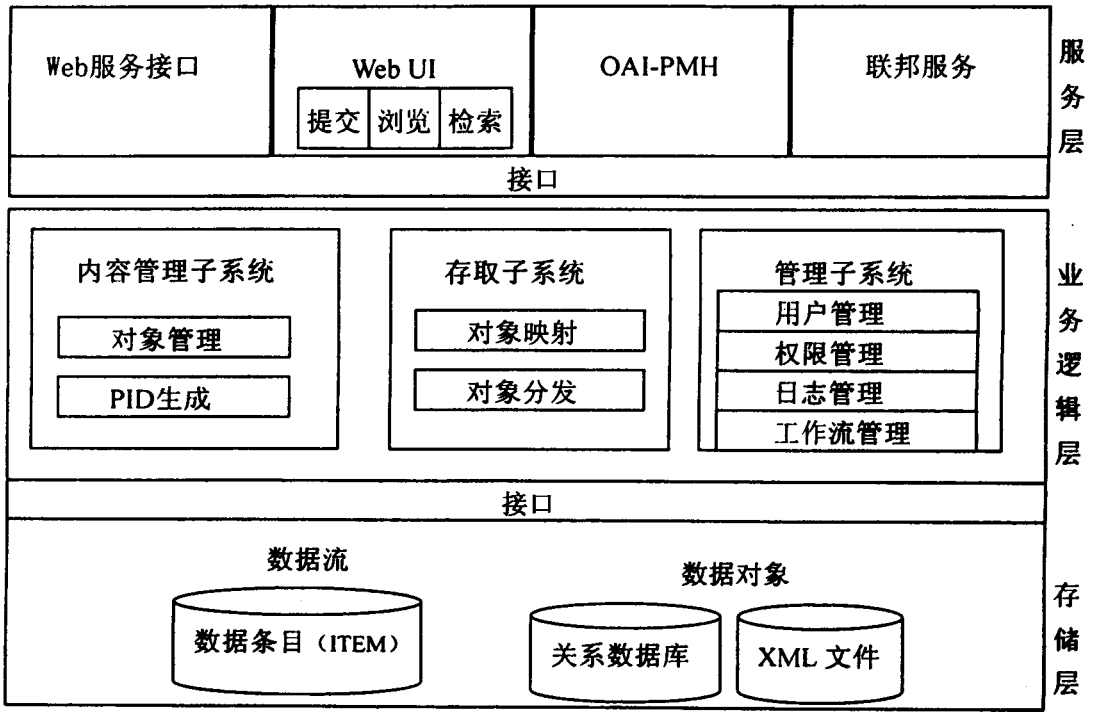

图1 IR的体系架构图

2.2 IR的主要功能

(1) 数字资源的提交和收集功能

一般的用户(如学生、教师或科研人员等)可以通过Web方式提交数字资源,包括数字对象的基本元数据信息和数字对象的内容信息,管理员可以通过Web方式管理所有的数字资源(collection)和团体(community)信息(如团体的名称、LOGO、描述、下属单位等),可以单条输入或批量导入数字资源。

(2) 数字资源的描述功能

主要采用元数据来描述数字资源,可以处理多媒体对象、复合文档、试验数据、学习课件等复杂数字对象的描述问题,支持DC、METS、MPEG DIDL、IMS等多种格式和标准。对于符合一定格式要求的数字资源,实现元数据的自动生成和创建,以便在一定程度上简化用户和管理员的操作。

(3) 基于工作流的资源审核与发布功能

对用户提交的数字资源或工作人员增加的数字资源,采用工作流的方式进行审理和发布,包括批准、退回、退修、发布等工作流程,提供灵活的、多步骤的基于角色的工作流机制,允许机构根据具体需求定制各种工作流。

(4) 索引功能

支持对元数据的各个属性的索引,提供数字对象的全文索引,支持METS格式的索引,可以使用数据库本身的索引机制,也可以采用如Lucene、Google等成熟搜索引擎中的索引机制。

(5) 检索功能

用户可以通过Web方式进行各种方式的检索和查询,如按照团体名称、指定的字段、数字资源的类型、关键字以及全文等方式进行检索。允许按照机构或数字资源类别等来浏览整个IR资源分布结构,支持特定数字对象或文档的访问控制,允许有权限的用户直接下载数字对象的内容信息。

(6) 互操作功能

采用诸如CNRI的handle系统之类的标准或通用唯一标识符来唯一地标识数字对象,支持多种元数据格式标准,如DC和METS等。支持OAI-PMH协议,提供元数据的采集(Harvesting)和联合检索(federated searching)功能。

(7) 管理功能

提供用户管理和权限管理(如提交、查看、编辑、发布、保密等),支持多种用户认证方式(如用户名/密码、IP控制、x.509、LDAP等),提供特定数字对象的限定性访问,能产生各种统计报表,提供访问日志和历史信息。

(8) 存储和保存功能

支持包括文本、图像、音频、视频在内的各种类

型数字对象的存储,支持如文章、预印本、工作论文、技术报告、会议论文、图书、学位论文、数据集、计算机程序、可视化仿真环境和模型等任意形式的数字资源的保存。可以处理数字对象的多个版本,能够以数字对象的固有面貌进行输出。采用 XML 方式或成熟的数据库来存储元数据,提供可靠的备份机制和灾难性恢复机制。能通过移植、镜像、恢复介质或其他方式支持长期保存。

2.3 IR 的关键技术和标准规范

IR 的主要技术是数字对象管理技术和开放存取技术。

数字对象管理技术是 IR 实现内容组织和长期保存的关键技术,其核心内容是数字对象框架。应用于 IR 数字对象框架具有以下两个特点,一是数字对象动态绑定了数据和对数据的操作,二是统一对待元数据和数据本身。数字对象框架分为两层,即数字对象和数字对象容器。数字对象是 IR 的核心概念,它是一个被唯一标识的网络实体,可以封装、描述数字资源,并且提供访问数字资源的机制,主要包括唯一标识符、系统元数据、数据流和数据发布器四个部分,数字对象可用 METS、IMS Content Packaging 或 MPEG-21 DIDL 等来进行描述。数字对象容器用来容纳数字对象,并提供对数字对象进行管理和访问的接口。在数字对象容器内,按层次结构来组织多个数字对象,相关的数字对象构成了一个资源集合(collection),机构内一个部门或实体的多个资源集合组成一个团体(community),一个机构的所有团体则组成一个完整的机构库。

开放存取技术是 IR 实现互操作和开放存取的关键技术,主要包括基于 OAI-PMH 的开放元数据互操作技术、基于 DOI 的永久性保存与利用技术、基于搜索引擎的开放存取技术和基于 Web Service 的开放存取技术等。通过这些技术可以从多个途径为机构内外的用户提供学术资源的浏览服务。此外,IR 还需要支持如 OAIS 参考模型、METS 元数据标准(或类似的标准如 MPEG-21 DIDL 等)、OpenURL、DOI 和 Web Service 等标准规范。

3 IR 的发展趋势

随着信息化、网络化社会的不断进步,IR 在理论和实践上都在不断地向前发展,总体来看,IR 有以下几个发展趋势。

3.1 建立 IR 联盟

从目前情况来看,由于人力、物力、财力等因素的制约,并非所有的学术机构都愿意建立独立的 IR,因而建立分布式的 IR 联盟是未来发展的一种趋势。另外,由于科学技术的不断进步,使得科学交流和技术合作的范围不断扩大,一项科研成果很少由某研究机构独立完成,往往是由几个学术组织共同开发的。再有,在信息时代,科研工作者的整个职业生涯往往服务于不同的学术机构。这种情况下,建立 IR 联盟能够促进跨机构的学术合作,较完整地反映科研人员的研究成果及机构的学术资源。

目前学术界对于"IR 联盟"并没有一个清晰的界定,广义上讲,它指学术机构通过合作的方式,将各自的资源库整合起来,统一提供数字化服务,支持不同 IR 之间的数据交换和共享,支持跨库无缝检索,同时各分布式 IR 拥有数据备份、保存及故障恢复能力。如 FAIR 项目由英国的 40 多个机构共同建设、Sherpa 项目有 20 多个英国大学参加。目前已有各种支持 IR 联盟的解决方案,如 LANL 仓储体系、DSpace、Fedora 等。

3.2 IR 范围扩展

IR 范围的扩展主要体现在两个方面。一方面,IR 在一定程度上扩展其存储资源的服务范围,不仅仅为科研人员,也为社会公众提供信息服务。当今社会,学术团体的影响力和作用已不仅仅局限于高等教育和科学领域,它逐步向社会的各个层面渗透。另一方面,公共图书馆、地方博物馆、各种学会、档案馆等文献机构可以独立或通过联合的方式建立社区资源库(Community,Repository),保存其数字化资源,并为公众提供资料检索、获取等服务。社区资源库可在公开出版发行论文资料、为公众提供有偿服务的同时,获得一定的经济收益。

3.3 机构、出版商和 OAI 服务商等之间广泛合作

目前,已经出现了 IR 与搜索引擎服务商之间的合作,IR 向搜索引擎开放其元数据,使搜索引擎能查询到 IR 中的资源。例如,在 Yahoo 中可以检索 OAIster 中的元数据,在 Google 中可以查询 Dspace 中的元数据。随着 IR 的逐步普及,还将出现更广泛的合作,如机构、出版商、OAI 服务商等之间的合作和互操作,以及与作者实现互动等。例如,在 Sherpa 项目中,Oxford 大学与 OUP 出版社之间建立了合作和

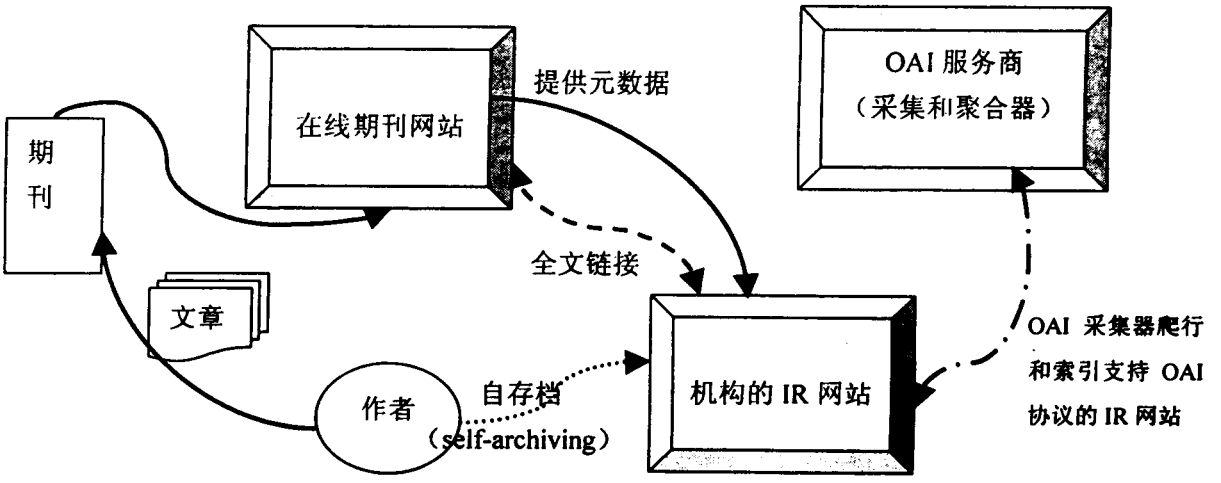

图2 IR与作者、出版商、OAI服务商的关系图

联系，同时还作为OAI数据提供商，支持OAI服务商（如OAIster）的元数据采集器。

图2示出了IR与作者、出版商、OAI服务商的合作关系。作者将自己的学术论文投稿给出版商，同时将论文提交给所在机构的IR。出版商向作者所属机构提供论文元数据，并给作者发送包含有论文全文链接的Email。在IR中可以检索到相关资源的元数据，并通过全文链接连到出版商的期刊网站。OAI服务商可以通过其采集和聚合器获取IR中的元数据，实现资源共享。

3.4 IR与其他系统和环境的无缝连接

随着IR实践活动增加，IR与机构内部以及机构外部资源和环境之间的集成和无缝连接也越来越受到关注和重视。在建设IR之前，机构通常已经建成了一些数字资源库，因此要考虑IR与现有系统以及资源库之间的融合。例如，Tufts大学的IR就实现与原有系统的无缝链接，通过接口使用原有系统发布TDL的数字资源，并且还根据仓储资源建立了VUE（Visual Understanding Environment）。

在与外部系统和环境的集成方面，利用标准协议来实现IR资源的开放服务，如Dspace支持SFX的OpenURL协议，利用DC元数据自动在每个Item页显示一个OpenURL链接。另外，利用Web Services技术来实现服务层次上的资源共享与互操作，也是近年来IR的一个重要发展趋势。Web Services具有完好的封装性、松散耦合、高度可集成性等优点，为实现IR与其他系统的集成提供了新的平台，如Fedaro框架可以通过SOAP来实现Web Services服务的互操作。

4 结束语

IR是在网络信息环境下，为解决学术交流体系中的矛盾而提出，它在一定程度上为学术机构解决所面临的经济、资源等问题提供了有效的办法，在一定程度上弥补了现有学术出版模式的不足，是对目前学术交流体系的一个重要的补充，推动了新型分离式学术出版模式的建立与发展。IR将会为学术交流领域人员，包括科研工作者、图书馆员、机构管理者等带来明显的效益，是应引起我们充分重视的研究领域。

参 考 文 献

1 Crow, Raym. The Case for Institutional Repositories: A SPARC Position Paper. SPARC: Scholarly Publishing & Academic Resources Coalition, 2002 http://www.arl.org/sparc/IR/ir.html[检索日期2005年4月8日]

2 Clifford A. Lynch. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL, 2003, (226):1~7

3 Mark Ware Consulting Ltd. Pathfinder Research on Web-based Repositories FINAL REPORT. http://www.palsgroup.org.uk/palsweb/palsweb.nsf/0/8c43ce800a9c67cd80256e370051e88a/$FILE/PALS report on Institutional Repositories.pdf [检索日期2005年4月8日]

4 Open Society Institute. A Guide to Institutional Repository Software v 3.0. http://www.soros.org/openaccess/software/[检索日期2005年4月8日]

5 Crow, Raym. Institutional Repository Checklist and Resource

Guide. SPARC: Scholarly Publishing & Academic Resources Coalition,2002 http://www.arl.org/sparc/IR/IR-Guide.html [检索日期 2005 年 4 月 8 日]

6 McCord, Alan. Institutional Repositories: Enhancing Teaching, Learning, and Research. Educause Evolving Technologies Committee, October 2003 http://sitemaker. umich. edu/dams/files/etcom-2003-repositories.pdf[检索日期 2005 年 4 月 8 日]

7 Johnson, Richard K. Institutional Repositories: Partnering with Enhance Scholarly Communication. D-Lib Magazine. November 2002 http://www. dlib. org/dlib/november02/johnson/11johnson. html[检索日期 2005 年 4 月 8 日]

8 Paul Wheatley. Institutional Repositories in the context of Digital Preservation. DPc Technology Watch Series Report, March 2004. httP://www.dpconline.org/docs/DPCTWf4word.pdf[检索日期 2005 年 4 月 8 日]

9 Nixon, William J. The evolution of an institutional e-prints archive at the University of Glasgow. Ariadne,2002,(32)

10 Pinfield, et al. Setting up an institutional e-print archive. Ariadne,2002,(31)

11 http://www.sfxit.com/OpenURL/[检索日期 2005 年 4 月 8 日]

12 http://www.bibl.ulaval.ca/archimede/index.en.html[检索日期 2005 年 4 月 8 日]

13 http://www.bepress.com/repositories.html[检索日期 2005 年 4 月 8 日]

14 http://cdsware.cern.ch[检索日期 2005 年 4 月 8 日]

15 http://contentdm.com/[检索日期 2005 年 4 月 8 日]

16 http://www.dspace.org[检索日期 2005 年 4 月 8 日]

17 http://software.eprints.org[检索日期 2005 年 4 月 8 日]

18 http://www.fedora.info/index.shtml[检索日期 2005 年 4 月 8 日]

(责任编辑 芮国章)