文章编号:1006-2467(2003)S1-0016-05

台湾大学典藏数字化计划——建置以OAI-PMH为基础的互通检索系统

陈雪华
(台湾大学 图书信息学系)

摘 要:OAI之互通性架构,在欧美等地已有许多机构与研究单位着手进行相关之研究与实际系统的建置与使用."台大典藏数字化计划"以OAI架构为其基础,进行与互通机制相关的研究与系统规划.文中介绍OAI概念架构,并就"台湾大学典藏数字化计划"之互通性检索系统规划与建置加以介绍及提出讨论.

关键词:互通性系统;数字典藏;数字典藏国家型科技计划;台湾大学典藏数字化计划;元资料
中图分类号:G250.7 **文献标识码**:A

Building an OAI-based Interoperable Retrieval System for the Taiwan University Digital Archives Project in Taiwan

CHEN Xue-hua
(Dept. of Library and Information Science, Taiwan Univ.)

Abstract: The interoperable framework of OAI has received much attention from scholars of library and information sciences. In North America and Europe, many academic organizations and universities have undertaken theoretical studies, system design, and implementation of the OAI framework. Taiwan University Digital Archives Project (TUDAP) is one of its institutional projects, proceeding relative studies and system design based on OAI framework. This article explained the reason why the OAI-PMH was chosen as the interoperable mechanism of the TUDAP, and introduced a prototype system of OAI-PMH implemented within this project.

Key words: interoperable system; digital archive; open archives initiative protocol for metadata harvesting (OAI-PMH); Taiwan university digital archives project (TUDAP); metadata

能够透通地(Transparency)获得所需的资源,而不需逐一到各个数字图书馆内搜寻,是多数检索者企盼许久的事;然而其中最主要的阻碍之一便是因为许多数字图书馆之间使用不同且特有的技术,因此,缺乏相互沟通的机制已经是现今数字图书馆之间所面临的重大问题之一[1].数字图书馆或学术机构的数据库与系统彼此互不隶属,相关数据或不同领域的数据不仅分散储存而且难以整合,使得数据的分享与流通有所限制.有鉴于此,1999年7月由Paul Ginsparg等[2]发出一份针对学术电子文件(e-print)分享计划的Universal Preprint Service会议邀请,于10月在美国新墨西哥州Santa Fe所举办的会议中推动了开放性数据库发展协会(Open Archives Initiative, OAI)的成立[3],并发表了OAI元资料撷取协议(Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMH),提供数字图书馆与博物馆检索互通(interoperability)机制上的另一可行方案.

OAI之互通性架构,在欧美等地已有许多机构

作者简介:陈雪华,教授.

与研究单位着手进行相关之研究与实际系统的建置与使用.在台湾地区,"台湾大学典藏数字化计"其下包含 7 个子计划;如何让各子计划所建置的系统得以分享、使用者借由单一接口即能检索到所有典藏单位的数据、以及让数字典藏的全貌得以展现,互通机制的建立为当务之急.为解决此问题,"台大典藏数字化计划"以 OAI 架构为其基础,进行与互通机制相关的研究与系统规划.本文介绍 OAI 概念架构,且就"台湾大学典藏数字化计划"之互通性检索系统规划与建置加以介绍并提出讨论.

1　OAI 元资料撷取协议的内涵

1.1　OAI 元资料撷取协议之制订缘起

OAI 协会最初是由 Paul Ginsparg, Rick Luce, Herbert Van de Sompel 等人,在 1999 年 10 月于 New Mexico 的 Santa Fe 所举行的 Universal Preprint Service 会议中所促成.由 Council on Library and Information Resources(CLIR)、Digital Library Federation(DLF)、The Scholarly Publishing & Academic Resources Coalition (SPARC)、The Association of Research Libraries(ARL)及 Los Alamos National Laboratory(LANL)等组织所赞助;执行委员为 Carl Lagoze 与 Herbert Van de Sompel 两位教授.与会者主要来自计算器科学及数字图书馆领域,就期刊论文及预刊本之电子资源互通性的议题进行讨论.有鉴于各数据库系统间,彼此互不隶属,相关数据、或不同领域的数据,分散储存而难以统整,使得数据的流通有所限制而未臻完善.会议中对于电子期刊预刊本(electronic pre-print)及其相关之学术论文数据库,与会代表形成必需发展出一套具互通性的标准技术架构的共识;透过技术机制及组织架构的发展以支持电子出版品档案(e-print archives)的互通性[3].

OAI 组织参考了由 C. Mic Bowman, Peter B. Danzig, Darren Hardy, Udi Manber, Michael Schwartz 等人所建置的 Harvest 系统架构[4],并经过先期版本测试一段时间后,于 2001 年 1 月 21 日正式公布了 OAI-PMH 元资料撷取协议[5],同年 7 月 2 日修订发表 1.1 版[6],最新版本为 2002 年 6 月 4 日正式发表的 OAI-PMH 2.0 版[7].OAI 已经将 OAI-PMH 提交 W3C 计划申请为国际标准,以期能成为全球元资料分享的开放性标准.借由此协议具备应用程序独立,且可互相运作的优点,以提供和提升 Web 上多种从事文件内容出版发行的社群应用与分享的服务功能.

1.2　内涵与规范

OAI-PMH 协议的内涵主要包含元资料的发布(expose)与撷取(harvest)两个类型,其在协议上分别定义此两部分:

(1)定义一个数据提供者能够透过 HTTP 为基础的协议,发布其元资料(metadata)的机制.

(2)定义一个能够从储存器(Repository)获取含有元资料数据录(record)的机制.

加入 OAI 的组织依据任务的不同,主要区分为数据提供者(Data Provider)与服务提供者(Service Provider)两个角色.加入者须依据本身服务提供的种类,向 OAI 登记注册成为数据提供者或服务提供者.OAI 在接受登记注册后,会执行相对的验证以确保登记之系统能够完全符合 OAI 协议的规范.除此之外,登记注册还有下列目的[8]:

(1) 能够提供 OAI 确认过之储存器(Repository)的明细,使服务提供者能够很容易的知道有哪些能够获取数据来源的储存器.

(2) 确保数据提供者能够提供符合 OAI 协议规范的机制.

(3) 提供 OAI 能够监督协议的使用,以及规划未来的活动与策略.

1.3　OAI 元资料撷取协议之技术架构

OAI-PMH 整体的运作如图 1 所示,核心主要是在 HTTP 协议上传输使用 XML 文件的协议,而前后端整个运作环境包含下列五个主要的组成组件[8]:

(1) 数据提供者.提供其文件内容,并以 OAI-PMH 作为发布元资料的协议.主要工作为维护一个或一个以上支持 OAI-PMH 来将其内容以元资料发布的储存器(Web 服务器).

(2) 服务提供者.透过 OAI-PMH 协议向数据提供者取得数据,并利用获得的元资料建立各种加值服务.

(3) 数据储存器.透过 HTTP,接受 OAI-PMH 所提出存取数据需求的服务器.

(4) 资料集.非必备功能,主要是为了方便取得部分范围所需的数据.储存器内可将不同类别的数据区分成不同的群组,并以阶层式架构表示,以节点(node)作为各分类的区分,因此每一个节点即称之为数据集.各数据集之下可再细分数据集,因此服务提供者向后端数据提供者获取数据时,便可指明特定数据集范围的数据.

(5) 资料录.一个数据录是后端服务器依据 OAI-PMH,从储存器内将数据以 XML 编码传回前端的元资料.数据录的元资料结构包含下列三个部分:

①标头信息(header):标记有关此笔数据纪录

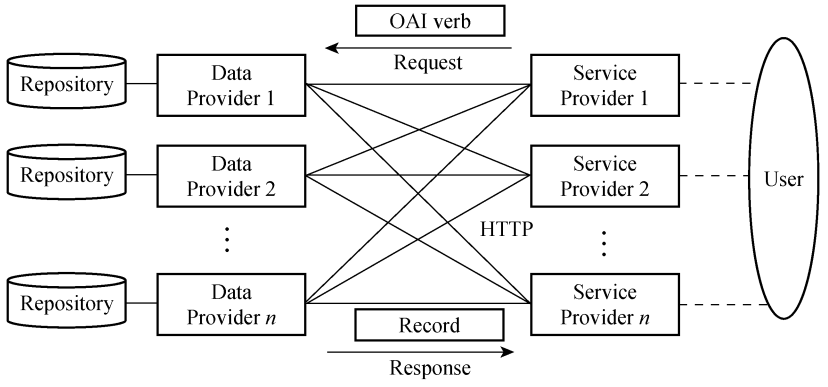

图1 OAI-PMH技术架构组成组件

的信息,分为两部分:一为专有识别名称(unique identifier),是数据库内数据独一无二的识别标志;二为日期戳印(datestamp),有关数据新增、维护、删除的日期信息,提供服务提供者端程序判断处理.

②元资料(metadata):数据录实际的元资料内容.OAI-PMH规定资料提供者必须支持Dublin Core,是否提供其他元资料格式则由数据提供者端的系统决定.

③相关信息(about):非必备的部分.提供有关资料的相关说明,如版权宣告、使用权限等,这部分的结构在OAI-PMH中不做规定.

1.4 OAI元资料撷取协议的优势[9]

OAI-PMH之互通性架构规范具有许多优点,分述如下:

(1) 提供学术沟通及交流的新模式. OAI-PMH之架构使数字化文件能更容易、更广泛的传播.且采用元资料撷取的方式,可涵盖各种多媒体格式、数据型态、与内容等,扩展了数字化数据存取种类的范围.

(2) 实作容易. OAI-PMH在设计时即以"简单"为原则.一个具架设网络服务器经验的工程师,可以在极短的时间内架设起OAI服务器.

(3) 具开放性(open). 任何人都能使用OAI-PMH定义的架构,来建构数据提供或服务提供的服务器.

(4) 采用HTTP及XML之开放性标准. OAI-PMH利用HTTP通讯协议作为其基本的通讯协议.其优点在于:现今所有的网页服务器及浏览程序等,几乎毫无例外的支持HTTP;这使得OAI在先天上就已解决了跨平台及兼容性等问题,也节省了另行架构的困难.另一方面,XML亦渐渐成为全球共同的数据标准交换格式.由于HTTP及XML均为开放性的标准,采用HTTP及XML的组合不仅考虑了兼容性的问题,也确保了OAI-PMH的开放性原则.

OAI已经将OAI-PMH提交W3C计划申请为国际标准,以期能成为全球元资料分享的开放性标准.借由此协议具备应用程序独立,且可互相运作的优点,以提供和提升Web上多种从事文件内容出版发行的社群应用与分享的服务功能.此一协议并不是要取代目前的Z39.50等开放式资料存取标准,而是提供一个容易实作(easy-to-implement)和容易部署(easy-to-deploy)的替代性解决方案[10].它的出现只是要让Internet网络上众多的信息服务提供者,以一组标准的陈述语法,可以让不同的服务提供者及资料提供者互相沟通没有组碍.此即为"台湾大学典藏数字化计划"将此一协定纳入研究与实作范围之内的主要原因.

2 "台湾大学典藏数字化计划"背景

OAI相关发展计划在世界各地如火如荼地展开,台湾地区亦有相关计划正在着手进行.2001年台湾地区图书信息相关技术规范汇编之"分散检索标准"中,除继续就Z39.50协议进行研究外,OAI协议也已经被列为分布式检索标准研订小组的研究重点之一[9].

"台大典藏数字化计划"包含7个子计划,分别为:台湾文献文物典藏数字化计划、台湾大学植物标本典藏数字化计划、台湾大学昆虫标本典藏数字化计划、台湾大学地质科学典藏数字化计划、台湾大学人类学系典藏文物数字化计划、台湾大学动物博物馆典藏数字化计划等6项内容发展子计划及提供技术服务的台湾大学数字典藏技术服务计划.有鉴于典藏单位众多且独立,为使各典藏单位所建置的系统得以分享,使用者借由单一接口即能检索到所有典藏单位的数据,并能让数字典藏的全貌得以展现;"台大典藏数字化计划"拟规划一"台湾大学数字典藏资源中心"(NTU Digital Archives Resource Center; DARC),以OAI架构为基础的互通检索系统规划为

其发展重点之一. 期望就 6 项内容发展子计划的数字化成果,并在考量未来典藏与使用之需求下,提供一整合性的管理与检索界面,以利于使用与维护.

3 "台湾大学典藏数字化计划"互通检索机制建置

开放式档案分享机制是迈向开放式数字图书馆应该具备的服务之一,而所有数字典藏单位之间也应该能共同分享彼此的资源,提供使用者单一且透明的信息取得管道. 本计划的互通基础即是应用 OAI 协议,使各个数据提供者与服务提供者之间的沟通更为容易,使得数字典藏的数据能够保有元资料的原始结构或 Dublin Core 格式,并透过标准且简单的程序达到分享、使用与加值,有助于使用者更方便地检索与获取网络资源,满足文献信息检索的需求.

3.1　系统实作

本计划依据 OAI-PMH 2.0 版,将计划内六典藏单位之数字化成果整合于大数字典藏资源中心,其整体系统架构环境如上图 2 所示. 根据 OAI 协议,共分为数据提供者(安装于各典藏单位的系统上)、服务提供者(资源中心之整合系统)和数据库系统三部分. 各部分所需执行之作业范围分述如下:

(1) 数据提供者. ①批次将典藏单位之元资料转换成 XML 格式,及对映成 Dublin Core 格式; ②依据前端服务提供者所传送之需求,剖析(parse)并依据 OAI 协议定义之 XML Schema 格式封装元资料并响应至前端.

(2) 服务提供者. ①将后端数据提供者响应之元资料,包括 Dublin Core 及特有元资料之数据储存,并予以记录相关属性,包括下载时间、原始系统编号、数据来源等; ②将下载之数据内容依据索引参数定义,建立查询功能所需之索引档案; ③使用者授权及系统管理接口; ④查询功能.

(3) 数据库. ①记录与管理搜集自各单位所发布之元资料,及低分辨率的数字物件; ②检索所需之相关文件资料; ③系统管理使用之参数文件数据; ④提供 OAI 应用上所需的元资料对映参数与所属的 XML Schema 定义.

综合上述,所分析的单元组件与所属功能,在系统开发之初的前置作业必须先执行下列工作:

(1) 收集各典藏单位数据提供之深度,包括各单位元资料之结构、语法及语意,元资料及数字对象之连结关系,允许提供数据的程度与范围.

(2) 整理各典藏元资料转换 Dublin Core 字段之对照表,并建立所属元数据 XML Schema 的定义.

(3) 各典藏单位之元资料域位属性的索引点(access point)定义.

(4) 建立各单位元资料之间相关字段的关联性,亦即制定元资料之规范文档控制.

3.2　实作过程探讨

本系统为依据 OAI-PMH 2.0 版,借由实作数据提供者及服务提供者两端的程序,作为台湾大学数字典藏资源中心系统的运作核心. 本系统采用 Java 作为系统软件之开发语言,而数据库系统在开发时采用 Microsoft 的 SQL Server2000,不过只要符合 SQL92 标准之关系型数据库系统均可. 综观实作过程的经验如下:

(1) 因各典藏单位元资料性质的差异,必须另提供一共同字段之精简版元资料,而最适当的共同字段合集便是 Dublin Core 元资料.

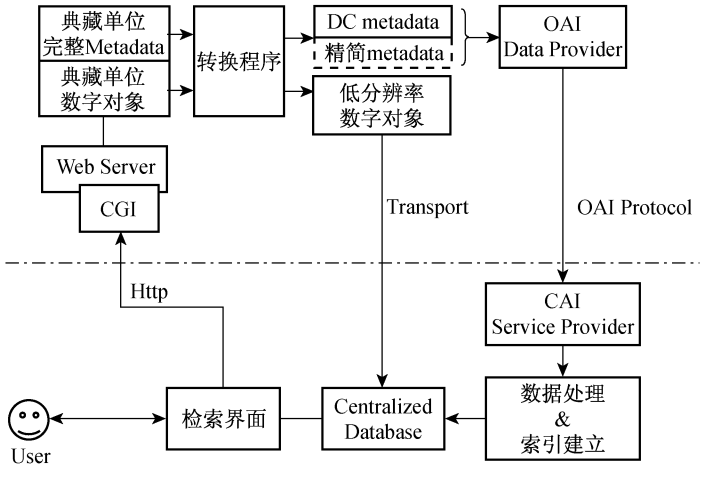

图 2　台大数字典藏资源中心架构图

(2)各典藏单位基于数据来源著作权等因素,部分数据并不适合公开,或只允许公开部分字段内容,因此在OAI Data Provider运作上必须先经资料过滤的处理过程.基于前项与本项需求,在各典藏单位除了需建置OAI相关之程序外,尚需包含数据异动与转换之处理程序.

(3)本质上,OAI只是一个用来交换元资料的协议,并不包含如文件、影像、声音等全文数据(full-content);其余文件格式、内容均需透过其他程序应用技术辅助,并不在此协议处理的范围.因此在整合资源中心的功能上必须要能连结回原典藏单位之数字对象所在.而在服务提供者端(资源中心),除了提供整合性的数据检索功能外,在呈现台大数字典藏资源中心架构图时,亦需显示简易的数字对象内容,例如缩图或部分多媒体档案内容.整体而言,服务提供者端除了利用OAI协议向数据提供者获取相关元资料外,亦需下载所需的简易数字对象数据.

4 结 语

开放式档案分享机制是迈向开放式数字图书馆应该具备的服务之一,而所有数字图书馆之间也应该能够建立联合目录,以便共同分享彼此的资源,提供使用者透通的信息取得管道.而OAI-PMH是发布和撷取元资料的开放式标准,借由此项标准可使各个数据提供者与服务提供者之间系统的沟通更为容易,使得数字典藏的数据能够保有元资料的原始结构或Dublin Core格式,并透过标准且简单的程序达到分享、使用与加值,有助于使用者更方便地检索与获取网络的资源,满足元信息检索的需求.经过初步的测试、评估后,本计划认为OAI-PMH的确具有简单且易建立的特性.预期本计划以OAI为基础所建置的系统,将能使得各典藏单位之间能够很容易地分享与整合彼此的元资料、数字对象;进而提供使用者正确且快速的信息服务,未来并可做为进一步与其他典藏单位或国外系统互通与合作的基础.

参考文献:

[1] Xiaomin Liu, *et al*. Arc ——an OAI service provider for digital library federation[J] **D-Lib Magazine**, 2001, 7(4). available at http://www.dlib.org/dlib/april01/liu/04liu.html

[2] Ginsparg Paul, Rick Luce, Herbert Van de Sompel. Call for participation in the UPS initiative aimed at the further promotion of author self-archived solutions. 1999, 7. http://www.openarchives.org/ups-invitation-ori.htm

[3] Herbert Van de Sompel, Carl Lagoze. The Santa Fe Convention of the Open Archives Initiative[J]. **D-Lib Magazine**, 2000, 6(2). http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

[4] C Mic Bowman, et al. The harvest information discovery and access system[J]. **Computer Networks and ISDN Systems**, 1995, 28(1-2): 119-125.

[5] The Open Archives Initiative Protocol for Metadata Harvesting, v.1.0, http://www.openarchives.org/OAI/1.0/openarchivesprotocol.htm

[6] The Open Archives Initiative Protocol for Metadata Harvesting, v.1.1, http://www.openarchives.org/OAI/1.1/openarchivesprotocol.htm

[7] The Open Archives Initiative Protocol for Metadata Harvesting, v.2.0 http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[8] 余显强, 陈瑞顺. 开放式档案存取协议——OAIMH[J]. **书艺**, 2002, 38(5): 25-39.

[9] 陈雪华,等. 数字典藏互通性架构之探讨[J]. **中国图书馆学会会报**, 2002, (6): 1-13.

[10] 余显强. 浅谈开放式元数据采集协议的应用与内涵[J]. **图书与信息学刊**, 2002, (11): 34-45.