文章编号：1005-8451（2006）02-0009-03

基于OAI协议的统一检索系统研究与实现

张 浩，黄厚宽
（北京交通大学 计算机与信息学院，北京 100044）

摘 要：OAI协议为实现一个统一、高效的统一检索系统创造了条件。基于一个数字图书馆项目，主要介绍OAI概念架构并阐述一个基于OAI协议的统一检索系统的设计与实现。

关键词：OAI；数字图书馆；统一检索；研究
中图分类号：TP39 **文献标识码**：A

Research and implementation of Unified Search System based on OAI protocol

ZHANG Hao, HUNG Hou-kuan
(School of Computer and Information Technology , Beijing Jiaotong University, Beijing 100044, China)

Abstract: The OAI protocol presented such a condition that it was able to design and implement a uniform, highly efficient Unified Search System, On basis of a Digital library project, it was introduced OAI framework, and elaborated a design and implemented for unified search system based on OAI protocol .

Key words: OAI; digital library; unified search; research

近年来，数字图书馆发展迅速，各个学校图书馆通过引进和自建数据库，已使电子资源的建设具有相当规模，电子文献在文献服务中所占的比重也不断增加。随着数字图书馆技术的不断发展，各种数字资源层出不穷，同时，由于数字资源建设的不同步以及采用技术的不同，各种数字资源都有自己的数据结构、组织方式、查询方式以及显示界面。对于用户来说，为了查准、查全所需要的资料，不得不分别进入不同的查询系统，熟悉每个数据源的检索方式和显示格式。

跨平台检索系统正式针对这个问题而出现。它可以在一个统一的界面和查询环境下对不同数据源的信息统一进行查询，并以统一的界面显示不同数据源的信息。跨平台检索系统可以节省用户获取资料的时间；提高查准率和查全率；将不同媒体不同类型的数据源以整合的方式显示。随着读者对这种需求的强烈要求，各个高校的数字图书馆开始更加关注电子资源的管理工作，整合已有的资源，将不同类型、不同结构、不同环境、不同用法的各种异构数据库纳入统一的检索平台，以便于用户更方便、更高效地获取信息。

现今跨平台检索系统有几种类型，都有各自的特点和适用范围。但是在结构和应用上尚不能达到真正的统一资源整合发布要求。近期，本人参加了某高校的数字图书馆建设及相关全文检索系统的开发任务，对如何构造一个基于“OAI-PMH”协议的跨平台跨媒体的统一检索系统进行研究和探索，本文是此项目研究的一个总结。

1 OAI技术介绍

1.1 OAI简介

OAI（获得元数据的开放信息仓库首创协议，Open Archives Initiative protocol for metadata harvesting）是一种利于有效的传播书目的技术，当前主要的应用是交互式的搜索信息系统。协议的目的是为了提供和促进互连网上发布内容的多个团体的与应用无关的交互操作。

1.2 OAI的组成

协议主要由两个方面的交互操作组成：（1）Data Provider administer system：发布元数据方。拥有信息仓库（Repository），发布数据，使得终端使用者或服务提供方可以使用浏览仓库；（2）Service Providers：服务提供方。向发布元数据方发出请求（requests），并接收返回的元数据作为构造附加服务的基础。OAI原理图如图1所示。

收稿日期：2005-08-15
作者简介：张 浩，在读硕士研究生；黄厚宽，教授。

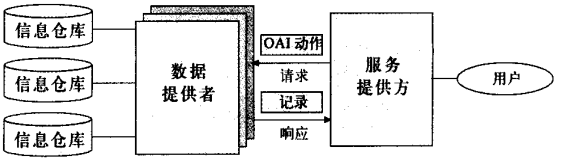

图1 OAI原理图

2 基于OAI协议的统一检索系统的实现

统一检索系统是迈向开放式数字图书馆应该具备的服务之一，而所有数字资源单位之间也应该能共同分享彼此的资源，提供给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，使得数字资源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。

2.1 统一检索系统整体框架

本计划依据OAI 2.0版，将某高校各个院系及图书馆之数字化成果整合于数字图书馆资源中心，其整体系统架构如图2所示。根据OAI协议，共分为数据提供者（安装于各院系单位的系统上）、服务提供者（数字图书馆资源中心之整合系统）和数据库系统3部分，各部分功能分述如下：

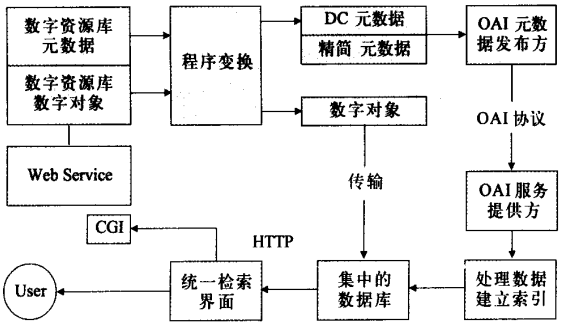

图2 统一检索系统整体架构图

（1）数据提供者：a.分批次将提供数字资源单位之元资料转换成XML格式，并且映成Dublin Core格式；b.依据前端服务提供者的需求，剖析并依据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：a.后端数据提供者的元资料，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；b.将下载之数据内容依据索引参数定义，建立查询功能所需的索引档案；c.用户授权及系统后台管理接口；d.检索功能。

（3）数据库：a.存储各院系所发布的元资料和低分辨率的数字图片；b.检索所需之相关文献资料；c.系统后台管理使用的参数数据；d.提供OAI应用上所需的元资料对映参数与所属的XML Schema定义。

2.2 基于OAI-PHM的论文网上提交系统

2.2.1 论文提交

（1）提交权限认证：要求系统提供两种选择，各校根据各院系情况自行选择：a.不需要权限认证；b.需要权限认证。

（2）提交表单项目：要求系统提供表单项目的配置，系统必须包括核心表单项目，其它表单项目可以由管理员根据本校情况，通过可视化界面自由配置。

（3）全文文件名

一般选择“学号+院系代码+论文全文格式.扩展名”组合来确定文件名，论文全文格式为word，需要区分2000、2003或XP，扩展名取自学生提交的文件名。

例如：03122119 -211030-word03.doc（表示学号为03122119提交的word2003的文件）；

（4）提交结果查询

通过输入认证条件，如学号+密码等来查询，是否提交成功，或者修改后再次提交。

2.2.2 管理员审核

（1）记录处理

对学生提交的记录逐条进行检查，包括检查论文文摘等元数据和论文全文。

不合格的论文给出不合格原因，系统最好提供常用的不合格信息列表，方便管理员选择。不合格信息通过两种方式返给学生：自动发E-mail；学生通过提交结果查询页面选择。系统还提供按院系分配任务功能，不同的管理员分别管理不同院系的论文。

（2）记录统计

可按院系统计，未处理的纪录、不合格的纪录和合格的纪录总数，统计结果可按照姓名、学号、提交日期、培养单位等项目排序。

（3）记录删除

（4）审查合格的记录立即发布

2.2.3　管理员编目

论文发布年限确定有两种方法：

（1）与学校论文主管部门协商，由学校统一规定论文的服务年限，例如，内部的论文统一规定为3年之后在因特网上发布，秘密的论文统一规定为8年之后发布；

（2）论文的发布年限由学生和导师共同确定，学生提交给图书馆的授权书上注明发布年限；针对这种情况，系统最好提供：批量加入年限；一条条地加入年限，系统读到年限后，自动将论文服务的权限放开。

2.2.4　文档标准化

（1）将学生提交的word文件转成pdf文件；

（2）自动的批量转换，不需要人工干预，系统分3个目录，“word文档”、“转换成功的pdf文档”、“转换不成功的pdf文件”，系统自动将转换的文件放入相应的目录中；

（3）在转换的同时，系统生成两个文件：完整的pdf文件；前24页的pdf文件。

本模块的流程图如图3所示：

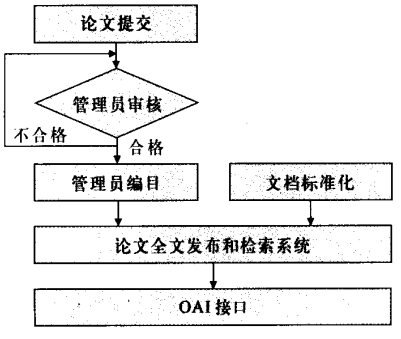

图3　基于OAI-PHM的论文网上提交系统流程图

2.3　基于OAI-PHM的全文发布和检索系统

（1）访问权限：基于用户和IP访问控制。

（2）检索：提供简单检索和组合检索功能，可进行二次查询，也可进行智能扩展检索；检索字段可由管理员来配置。一般检索字段包括题名、作者、导师、文摘和全面检索等；检索词之间可进行逻辑组配；提供按学科分类浏览功能。

（3）论文统计管理：提供对单篇论文浏览的总次数统计；根据IP地址范围对来访院校进行统计排名；浏览次数前30位论文的排名。

本模块的流程图如图4所示。

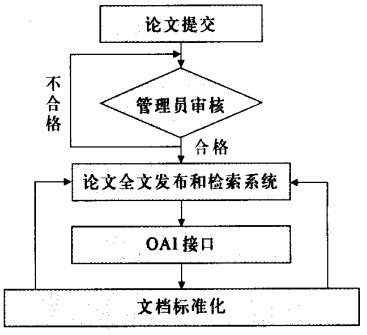

图4　基于OAI-PHM的全文发布和检索系统流程图

2.4　OAI接口

（1）支持OAI协议；

（2）能响应服务提供方的请求，并向之提供元数据；

（3）数字资源唯一标识符identifier统一规定为“学校代码+学号”。

3　结束语

统一检索系统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建立统一的原数据标准，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OAI2PMH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，使得数字资源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。经过初步的测试和评估后，本计划认为OAI2PMH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，将能使得各数字资源提供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，未来可作为进一步与其它数字资源单位或国内其它大学数字图书馆互通与合作的基础。

参考文献：

[1] James K，SEBENIUS.Negotiation Analysis:A Characterization And Review[J].Management Science，1992，38（1）.

[2] 王爱华，张　铭. 基于OAI的数字图书馆中元数据互操作框架[J]. 计算机工程与应用，2002（1）：5—7.