Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

统一检索技术在dILAS平台中的实现与应用

□ 秦格辉 / 深圳图书馆 深圳 518036

摘要：异构数字资源服务是当前数字图书馆工作中一个重要的课题，如何建立数字资源统一检索平台，为读者提供简捷、高效的检索服务显然是重中之重。文章试图从应用实际需求出发，通过记录多种成功技术实践，在分析比较目前几种流行的统一检索平台技术的基础上，以国家重点数字图书馆科研项目“dILAS”在深圳图书馆业务及粤港澳书目互查等技术实现为例，详述其统一检索平台的实现机制、关键技术和实现过程，更以相当篇幅介绍了平台的构成、特点和具体技术实现方法。

关键词：资源整合，统一检索，Z39.50协议，HTTP协议，OpenURL
DOI：10.3722/j.issn.1673—2286.2009.04.012

1 引言

随着信息技术的发展，数字资源的建设与服务在图书馆起着举足轻重的作用。基于多种平台、结构各异的各种类型的数字资源成为图书馆的重要收藏源。在多种异构资源并存的情况下，如何为读者提供便捷、有效的文献信息服务，是每个图书馆必须解决的问题。

dILAS系统的统一检索平台旨在探索实用化的图书馆资源整合解决方案，包括图书馆的馆藏资源、联合目录资源、自建专题资源以及从数据库商购买的各种数字资源；建立异构系统统一检索平台，在多种不同的图书馆应用系统的基础上形成统一的服务体系。平台采用统一的检索界面和检索语言，除提供馆藏目录、目次、文摘、全文、图片等信息的检索外，还应与图书馆的各类应用系统相结合，如OPAC系统、馆际互借系统、原文传递服务系统、参考咨询系统、统一认证系统，为读者提供更便捷、更贴切、更人性化的服务。

2 统一检索平台的几种实现方法

目前，在解决异构数据库统一检索方面，通常采用以下几种方法：

（1）通过数据库接口软件与不同的数据库直接连接，如ODBC和JDBC等。在同时检索的数据库数量较少时，使用此技术可在一定程度上解决异构检索问题，但数据库达到一定数量时，处理速度很难保证。[1]

（2）通过对元数据的收集整理，安装在本地系统中，形成本地的大型元数据仓储。这种方式的优点是，经过收集转换后的元数据不仅格式统一，而且结构清晰，可以按照需求建立各种分类体系，或者按照更高级的知识本体对数据进行再组织和管理。但缺点是对于类型不同、分布广泛、更新频繁的数字资源，很难做到即时、准确地将数据收集齐全。

（3）运用元搜索引擎的基本原理，利用数据库的Web客户端进行统一检索。这种方法的缺点在于需要对各个数据库的Web处理接口进行详尽分析，若其中某个数据库的Web处理接口发生改变则需重新设计，接口的稳定性较差。[2]

（4）利用专业的检索协议，如 Z39.50协议，对异构数据库进行统一检索，这种技术的优点是能提供读者完整的目录资料，检索接口稳定，能快速实现资料传输，但缺点是要求掌握复杂的专业检索协议，而且要求所检索的资源必须提供对应的标准检索协议服务。

现有的大部分跨库检索系统都是以这四种方法为基础设计的，每种技术都有自己的优势和局限性。根据图书馆资源的内容特性和发布特性，单纯选用其中一种跨库检索技术是不能完全满足资源的整合服务需求的，必须结合多种检索技术。对于具有 Z39.50服务的数字资源，如图书馆馆藏资源、自建数字资源、联合目录资源等，都提供了标准的 Z39.50服务，因此优先采用Z39.50网关整合检索技术；而对于那些仅提供Web检索服务的网络数据库，则采用基于 HTTP协议的

Web浏览器技术。通过这两种技术的紧密结合，基本上可以解决图书馆数字资源的整合检索问题。

3 统一检索平台的技术实现

在dILAS统一检索平台中，选择了两种统一检索技术，构建了基于Z39.50协议和基于HTTP协议的统一检索平台。该平台基于Unix/Linux/AIX/Solaris操作系统，通过一系列资源配置参数，采用URL和Web页面分析技术，对各种电子数据库及专业网络数据库进行综合分析、统一控制，实现多个Web服务器和Z39.50服务器的广播检索[3]。

dILAS统一检平台不仅能支持多种元数据格式（CNMARC、USMARC、DC等）的检索，而且支持网页Form的检索，支持多种用户验证方式（用户登录、IP控制）及Cookie机制，支持多种字符集的互转（UNICODE、GBK、BIG5、CCCII）技术。以下将对平台的构成及具体实现技术作详细介绍。

3.1 资源配置

对于统一检索平台来说，关键问题是怎么整合千差万别的数据库，实现检索界面、检索结果的统一化。因此，首要的工作是要对各数据库进行分析，找出其检索流程、参数传递方式、结果提交及数据组织方式，我们可以通过多种工具（如 Visual Sniffer）来监控端口传输的数据流来完成这一步的工作。分析完各个数据后，得到它们的共同特性及各数据库的个性化特点，共性的部分可以通过固化的程序流程来实现，个性化差异作为参数，在资源配置中进行指定。当然，并不是每个数据库的所有检索方式都能通过参数配置来实现，只能做到求大同存小异，尽可能实现更多的数据库检索功能。

对所有数据库来说，其检索方式大都有以下特点：（1）用户认证->认证结果返回；（2）提交检索请求->返回检索结果列表；（3）提交详细数据请求->返回详细数据；（4）提交原文下载请求->返回原文。

因此在资源配置参数中，只要配置好数据库的授权级别、授权访问方式、检索方式（即索引转换表）、登录脚本、检索脚本、详细数据提取脚本等参数，就可通过一个统一的检索网关实现不同数据库的统一检索。

资源配置参数分两种，一为总控参数，其中包括平台控制参数和数据库连接参数。另一为针对数据库个性化检索而配置的检索脚本参数。通过这两类参数就可灵活增减检索数据库的种类及个数，而不需增加检索浏览器，更不需要对程序作任何修改。

（1）检索平台控制参数

[基本配置参数]

连接上限(MAX_SESSION)、超时处理机制SESSION_TIMEOUT、SERVER_TIMEOUT)、模版页面文件(START_HTML、RESULT_HTML)、服务器路径参数（SERVER_PATH、SCRIPT_PATH、LOG_PATH）

[检索服务器参数]

数据库类型(HTTP_SERVER、Z_SERVER)、服务器端口(SERVER_PORT)、数据库类型与检索浏览器对应参数(如CNKI=USPBrowser4CNKI、VIP=USPBrowser4VIP、BIB=ZCon等)

[语种分类参数(LANG)]

中文、英文

[学科分类参数(SUBJECT)]

法律、经济、教育、政治、综合、医学、电子通信、化学、生物学、材料学

[数据库类型参数(DBTYPE)]

电子期刊、电子新闻、会议论文、学位论文、专利、行业数据与报告、科技成果、电子图书、企业名录、法律法规

（2）数据库连接参数

[RESOURCE:#资源ID]

资源名称(NAME)、来源(SOURCE)、语种(LANG)、数据库名称(DBNAME)、数据库标识(DBID)、URL、字符集(CHARSET)、数据库分类、学科分类、统一检索服务器类型(USPSERVER_TYPE)、统一检索服务器IP、网络数据库检索脚本参数入口(SCRIPT_FILE)、原文索取请求连接(REFER_LINK)。

如果统一检索服务器类型为Z39.50服务器，则资源来源类型统一指定为BIB，在连接参数中还需指定Z39.50服务的IP、端口及授权用户(USERID、PASSWORD)、图书馆OPAC的入口(OPAC_URL)、详细数据模版页面(DETAILBIB_HTML)。

如果统一检索服务器类型为Web服务器，则通过网络数据库检索脚本参数(SCRIPT_FILE)配置各数据库的个性化参数。

（3）网络数据库检索脚本

[基本参数]

每页数据量(PAGE_SIZE)、是否代理全文数据下载(FULLDOWNLOAD_AGENT)

[授权级别和授权访问方式]

包括三级授权级别：目录级、详细数据（摘要）级、全文级。

授权方式有：AUTH: 0 完全限制、1 不限制、2 IP限制、3 用户限制、4 IP或用户限制。

被限制的IP(UNAUTH_IP)和被授权的IP(AUTH_IP)范围

[索引映射参数INDEX_MAP]

统一检索平台索引与当前数据库索引对照表

[登录脚本LOGIN]

数据提交参数(METHOD、ACTION)、登录结果匹配关键词(LOGIN_MATCH_STRING)

[检索脚本SEARCH]

数据提交参数(METHOD、ACTION)、检索结果集数据匹配模版(HIT_MATCH_STRING)、检索变量(CGIVAR)

[目录数据提取脚本PRESENT]

数据提交参数(METHOD、ACTION)、翻页变量类型（GOPAGE_VARTYPE=页码或记录偏移量）、翻页变量名称（GOPAGE_VARNAME）、检索变量(CGIVAR)。

[详细数据提取脚本FULL_PRESENT]

数据提交参数(METHOD、ACTION)

[原文下载脚本FULLDOWNLOAD]

数据提交参数(METHOD、ACTION)

3.2 平台构成

统一检索平台分为三层：用户接口层、服务分发层、协议转换层。

（1）用户接口层

用户接口层包括两部分：平台门户USPStart和统一检索入口USPGate。

USPStart为平台的入口，可按语种、学科或数据库类型等对资源进行分类显示，是图书馆电子资源及联合书目资源的门户。

USPGate为平台的检索接口部分，通过网页Form与用户进行交互，同时将检索请求传给后台浏览服务器。每当USPgate启动后，先向服务分发器USPServer发送登录请求，登录成功后, 服务器为之分配连接事务标识（SESSIONID）并返回为之服务的浏览服务器进程（SERVER_PID），随后USPgate直接与该服务器进行会话，将检索或数据提取请求发给它并等待它的处理结果。当接收到浏览服务器的处理结果后，USPGate直接将结果（已格式化的HTML文件）回传给客户。一次请求操作完成后，USPGate自动退出，下次请求将通过SessionID和Server_PID与对应的检索服务器交互。

（2）服务分发层

USPServer为平台的服务分发器，也是登录服务器。它接收所有来自客户端的登录请求，根据检索数据库类型及请求中的会话ID，分配相应的检索服务器。对于已登录过的会话，则直接分配前次为之服务的检索服务器。而对于新的会话，则先对所有的空闲检索服务器的负载情况进行综合分析，决定启动一个新的检索服务器或沿用一个空闲的检索服务器。任务分发完成后将对应的服务器PID传给客户端USPgate。如果所连数据库类型为Z39.50服务器，则直接将请求转发给Z39.50服务网关Zcon。[3]

（3）协议转换层

Zcon为平台与远程Z39.50的连接服务器, 也是远程Z39.50服务的客户端，通过它搭建起客户端与远程Z39.50服务器之间连接的桥梁。该连接服务由客户端发过来的“初始化消息”激活启动。每启动一个新的连接服务进程，就与远程Z39.50服务器上的Zserver建立连接，该连接一直保持激活状态，直到客户发来中断请求为止。连接服务启动后，将等待从客户发来的后续操作请求，并将其转化为Z39.50协议要求的数据形式，传给远程的Z39.50服务器；当连接服务接收到Z39.50服务的回应消息后，根据平台要求，形成需要的结果文件，回传给对应的客户端。

USPBrowser4***为平台的服务端, 也是远程网络数据库的检索客户端。该应用基于HTTP 协议，采用URL和Web页面分析技术，模拟人工检索方式，监听通讯端口，截取数据通讯包，获得检索过程数据和结果数据，根据检索脚本中的设定，对网页进行过滤，提取有用的数据信息，形成属于自己风格的新页面，提交给平台客户端USPgate。该服务器的服务终止方式与Z39.50网关服务的终止方式不相同，它没有具体的服务中断协议，只能通过超时机制来实现，当一段时间内都没接到服务请求时，服务器自动退出。

3.3 工作原理

通过统一的检索界面（USPGa t e）接收用户检索请求，服务程序（USPServer）根据数据库类型，将请求分发到为数据库定制的统一检索浏览器(USPBrowser4***）或Z39.50网关，各种浏览器根据自己所负责数据库的检索特点，转换检索请求，提交给各数据库的检索引擎，然后等待数据库的返回结果。当接收到检索结果时，对结果进行解析，提取其中的数据信息，重新组装变为统一格式发布。如图1所示，描述了在统一检索界面下对几种网络数据库、图书馆书目数据库等进行统一检索的工作原理及实现过程。

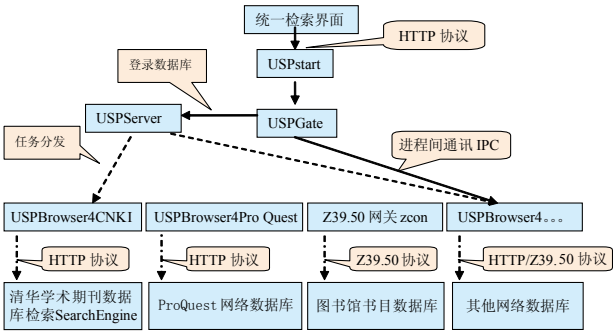

图1 基于HTTP/Z39.50协议检索平台工作原理

3.4 平台特点

● 简单的资源配置管理方式

通过在资源配置参数中设置的各种分类栏目（电子资源的多角度标引），可动态形成数据库的多种分类浏览页面，如按语种、学科、数据库类型等对检索数据库进行浏览检索。

● 连续的事务处理方式及严密的并发控制技术

平台基于UNIX操作系统，各个模块之间的信息交互及一致性控制采用了IPC通讯方式，即通过共享内存、消息队列、信号灯联合控制方法。共享内存记录当前活动的所有连接信息及资源使用情况，以保证HTTP请求的连续性。消息队列用来实现平台的客户端USPgate与连接服务端USPBrowser4*** / Zcon之间的请求交互。信号灯用来控制多进程间的互斥操作。

● 实时的馆藏链接服务

对于图书馆书目资源，可通过资源配置参数中的OPAC入口、详细数据模版页面及针对各馆配置的JavaScript脚本，动态生成书目的馆藏链接点，实时揭示文献的在馆情况。

● 简便的文献利用服务

将实体馆藏信息展示给读者的同时，结合馆际互借（ILL）协议，提供文献借阅请求登记服务。对于电子文献，则通过文件传输协议FTP、E-Mail等，直接将电子原文传递到读者手中，实现原文传递服务。对于没有开放授权的数据库，通过在检索结果页面中链接的原文索取登记页面，直接提交原文索取请求，后台工作人员代查后，将原文传给读者。

● 多种用户验证方式及Cookie机制

平台支持多种用户验证方式（用户登录、IP控制）及Cookie机制。

访问授权和版权控制，考虑到读者访问的方便和版权控制问题，实现了访问读者统一认证、单点登录方式。针对各电子资源的授权情况，进行分级服务，设置三级访问限定：元数据目录级、摘要级、全文级。访问授权方式分四种：0 完全限制；1 不限制；2 IP限制；3 用户限制；4 IP或用户限制。这样在方便读者服务的同时，也充分保护了电子资源供应商的版权。

● 支持多种字符集的互转技术

针对网上资源的情况，采用多种字符集的互转技术，包括UNICODE、GBK、BIG5、CCCII等，自动实现不同字符集之间的互检。

● 资源链接服务方便

统一检索平台利用各种资源定位协议（如HTTP、OpenURL、DOI等），在授权允许的情况下，对于提供开放式链接的电子资源，在展示元数据的同时，提供原文链接点，通过OpenURL技术直接链接到具体的全文数据或其他原始对象，方便读者联机获取。

4 统一检索技术的应用

结合图书馆运行的各类应用系统，搭建起图书馆统一检索体系，并总结出图书馆针对各种不同类型资源进行统一服务的过程，如图2所示。

图书馆统一检索体系首先在深圳图书馆付诸实施。深圳图书馆的资源结构复杂多样，既有自建的馆藏文献数据库、专题文献数据库，还有合作建设的数据库（如地方版联合编目数据库），购买的商用电子数据库，共享工程下的图书馆联盟数据库等。针对这些资源的特性和具体服务需求，搭建起多个统一检索平台，包括粤港澳书目检索、深圳市公共图书馆“通借通还”平台、深圳图书馆电子资源检索系统。

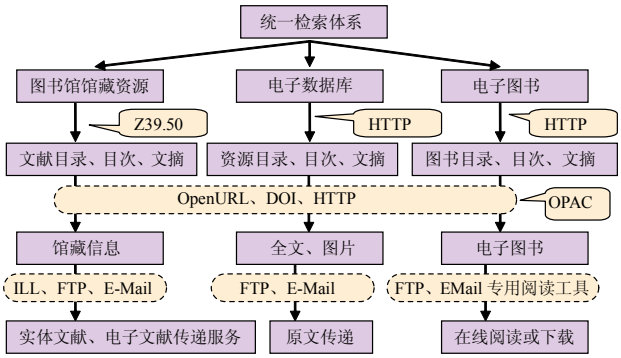

图2 基于多种检索技术的图书馆统一检索体系

4.1 粤港澳书目统一检索平台

该平台的建设是为了解决粤港澳地区图书文献资源的共享而建立的。实现了深圳图书馆、香港公共图书馆、澳门中央图书馆、省立中山图书馆和澳门大学图书馆的书目文献的统一检索。该平台基于 Z39.50检索协议，同时通过URL连接分析技术与各馆的OPAC系统结合起来，实现了粤港澳实时馆藏链接服务，为将来粤港澳图书馆进一步的资源共享、馆际互借打下了基础。同时该平台充分利用字符集互转技术，成功实现了Unicode、GBK、BIG5、CCCII几种字符集之间的互转，并实现了简繁通检。

4.2 深圳市公共图书馆图书“通借通还”平台

基于 Z39.50检索协议和馆际互借协议（ILL），实现深圳市馆与6个区馆及3个社区馆之间的 “通借通还”，系统在统一检索平台的基础上，与各馆的OPAC系统相连，能实时查看文献的在馆情况，在此基础上实现网上预借和馆际互借的功能。[4]

4.3 深圳图书馆电子资源统一检索平台

基于 HTTP 协议，采用URL和Web页面分析技术，用于对图书馆购买的各种电子数据库及其他专业网络数据库进行统一检索。目前实现统一检索的有50多种中外文数据库，如CNKI的中国学术期刊、博硕士学位论文、报纸全文数据库，重庆维普数据库、万方数据库、EBSCO、ProQuest、FirstSearch、Inspec等类型数据库。

5 小结

目前深圳图书馆各种类型的资源检索服务平台均已投入使用，特别是电子资源统一检索平台，深受读者的欢迎。但也必须看到，资源整合和统一检索服务是一项长期、复杂而烦琐的工作，要不断追踪新技术，不断跟进资源的变化，不断增加新引进的资源，不断将统一服务推向深入。

我们将继续加强统一检索平台建设，扩展服务范围，与图书馆各种服务系统紧密结合，包括馆际互借系统、原文传递服务系统、参考咨询系统、网上书评及推荐系统等，为读者提供更深层次的信息服务。

参考文献

[1] 黄锵.异构数据库的跨库检索技术综述[J].图书情报工作,2003(6):94-97,109.
[2] 王亮,郭一平.基于Webservice 的异构数据库检索系统[J].大学图书馆学报,2004(1):29-31,67.
[3] 秦格辉.Z39.50技术在ILAS系统中的应用[J].现代图书情报技术:2000(5).
[4] 余胜.深圳市公共图书馆图书“通借通还”全面开通[J].中国图书馆学报,2006(2):65.

作者简介

秦格辉，女，深圳图书馆副研究馆员。研究方向为图书馆自动化及数字图书馆技术。通讯地址：深圳图书馆 518036。E-mail: qin@szlib.gov.cn

The Application of Union Search in the System of dILAS

Qin Gehui / Shenzhen Library, Shenzhen, 518036

Abstract: Heterostructure data service is one of the most important project in current digital library research, and the focus of which is how to build an united search platform (USP) so that readers can search the data conveniently and high-efficiently. Based on the practical demands and the successful technical practices, the paper expatiates on the application schema, key techniques and the implement procedure of USP service in Shenzhen Library, part of dILAS which is a national key project of digital library research. Later it dilates on structure, characteristics and the implement methods of USP service.

Keywords: Resources integrating, Union search, Z39.50 protocol, HTTP protocol, OpenURL

（收稿日期：2009-02-15；责任编辑：贾延霞）