CADAL数字资源整合检索研究
——以清华大学图书馆OPAC系统为例*

远红亮，张蓓，张成昱，周虹
（清华大学图书馆，北京 100084）

摘要：清华大学图书馆OPAC系统利用CADAL元数据开放接口整合检索CADAL数字资源，在页面呈现检索命中记录的全文访问链接，以方便读者获取相关资源，提高CADAL数字资源呈现和利用率。本文着重描述清华大学图书馆OPAC系统整合检索CADAL数字资源的设计和实现方案，以供同类型图书馆和系统设计人员参考借鉴。

关键词：OPAC；CADAL；元数据；开放接口
中图分类号：G250.7
DOI：10.3772/j.issn.1673-2286.2017.08.005

1 引言

越来越多的读者在信息资源查找时首选网络搜索引擎而非图书馆OPAC（Open Public Access Catalogue），一是由于互联网技术不断发展以及互联网内容越来越丰富等情况改变了读者查询资源的习惯，二是OPAC系统功能不能满足读者日益增长的应用需求。因此，近年来图书馆工作人员不断尝试各种混搭应用，以期为读者提供更加多样化、丰富的功能选择[1]。清华大学图书馆将自身资源与大学数字图书馆国际合作计划（China Academic Digital Associative Library，CADAL）相关资源进行整合，使读者可在清华大学图书馆OPAC系统中检索到本馆资源以及CADAL数字图书馆的相关资源，从而提升读者检索体验。

CADAL作为国家教育部公共服务体系建设的重要组成部分，在资源、服务和技术等方面构成我国高等教育数字图书馆的重要基础。经过长期努力，CADAL已拥有多学科、多类型、多语种的海量数字资源，由国内外的图书馆、学术组织、学科专业人员广泛参与建设与服务，是具有高技术水平的学术数字图书馆，成为国家创新体系信息基础设施之一。CADAL以100万册（件）数字资源为核心，构建了由2个数字图书馆技术中心（浙江大学、中国科学院研究生院）和14个数字资源中心（北京大学、清华大学、浙江大学、复旦大学、南京大学、中国科学院研究生院、上海交通大学、西安交通大学、武汉大学、华中科技大学、吉林大学、中山大学、四川大学、北京师范大学）组成的分布式资源建设、组织和传播体系[2]。截至2015年12月31日，CADAL数字图书馆资源入库量达2 757 413册（件），在线量为2 428 656册（件）[3]。

设计与实现清华大学图书馆OPAC系统整合检索CADAL数字资源的功能，能提高CADAL数字资源的揭示、呈现和利用率，优化读者使用OPAC的检索体验，便于读者获取所需资源全文。本文旨在对清华大学图书馆OPAC系统整合检索CADAL数字资源的设计、实现和应用进行分析和研究，以期为同行提供参考。

2 分析与设计

2.1 需求分析

清华大学图书馆OPAC系统采用美国INNOVATIVE公司的INNOPAC/Millennium系统，其与CADAL数字资源平台完全独立异构。清华大学图书馆OPAC系统主要包括馆藏纸本图书和期刊、部分电子书刊和本校

* 本研究得到CADAL应用系统建设子项目“基于OpenAPI的信息检索系统”资助。

学位论文等；电子书刊主要以现代图书为主，来源于超星、书生之家和方正Apabi等数字资源库。CADAL数字资源平台涵盖古籍、民国书刊、现代图书、学位论文、其他多媒体资源，是对OPAC系统资源的有益补充。清华大学OPAC系统主要服务本校师生，管理本校数字资源。

CADAL数字资源平台依据内容版权要求提供电子版全文借阅服务，无版权限制内容面向全球用户服务，有版权限制内容仅面向合作单位用户服务。但读者只能在OPAC系统、CADAL数字资源平台分别进行检索，以获取所需资源。若要一次性检索出清华大学图书馆与CADAL数字图书馆中的书目信息，避免读者重复同样的操作，两个异构系统间必须进行数据交互。OPAC系统在检索本地数据的同时，对CADAL数字资源也进行相应检索，并将两部分检索结果整合后呈现给读者。若异构系统间没有开放接口进行元数据交互，页面分析法是一种解决方案；但这种方案较烦琐，若被分析页面结构发生变化，相应处理程序必须做出修改；不但影响用户使用，还可能导致系统故障。

异构系统间数据交互最好通过开放标准接口实现；清华大学图书馆OPAC系统没有标准的接口可供使用，需通过客户端脚本语言JavaScript编写程序进行页面处理。CADAL元数据开放接口是一套标准化通用接口，服务器端和客户端的编程语言均可以调用接口，方便使用JavaScript语言实现接口调用；CADAL开放接口系统中的元数据与CADAL数字资源平台中的元数据保持一致，通过元数据可组合出CADAL数字资源的全文访问路径，通过CADAL元数据开放接口系统可解决OPAC系统与CADAL数字资源平台数据交互问题；接口系统独立于CADAL数字资源平台，在整合检索过程中不会影响CADAL数字资源平台自身的访问，不会增加无谓的访问流量。基于上述分析，CADAL数字资源开放接口是连接图书馆OPAC系统和CADAL数字资源平台的桥梁。清华大学图书馆OPAC系统与CADAL数字资源整合检索功能，可通过调用CADAL数字资源开放接口实现。

2.2 接口分析

开放接口是应用编程接口（Application Programming Interface，API），其基于HTTP协议，以XML或JSON等格式返回数据和信息，主要为异构系统间数据资源互换和互操作提供方便。通过开放接口的方式，可实现信息和应用的关联和再加工[4]。CADAL元数据开放接口系统在图书馆OPAC系统和CADAL数字资源平台间架起桥梁，既独立于图书馆OPAC系统，又独立于CADAL数字资源平台。CADAL元数据开放接口系统与CADAL数字资源平台保持元数据实时同步，从而确保图书馆OPAC系统通过CADAL元数据开放接口系统获得的元数据信息准确无误，同时生成CADAL数字资源平台全文访问链接。截至2017年5月31日，CADAL元数据开放接口系统中元数据条目为2 433 880条，主要为古籍、民国书刊、现代图书、学位论文、英文图书等。

2.2.1 简单检索接口

CADAL元数据开放数据接口平台提供规范的调用接口，可实现CADAL资源元数据检索功能。以CADAL资源检索接口为例，采用发送HTTP请求的方式实现接口调用，基本形式为“http://IP或者域名/cadal/cbook/?q=检索字符串”。其中，“检索字符串=检索项:检索词”。检索项包括16个字段：BookNo（CADAL资源标识ID）、BookType（资源类型）、CreateDate（创建日期）、Creator（作者）、Publisher（出版机构）、Subject（关键词）、Coverage（覆盖范围）、Contributor（其他责任者）、ContentLanguage（内容语种）、Relation（关联）、Rights（版权）、Source（来源）、Title（题名）、Description（描述）、ISBN（国际标准书号）、Format（格式）。如“http://IP或者域名/cadal/cbook/?q=BookNo:xxxx”用于定位唯一CADAL资源记录，“http://IP或者域名/cadal/cbook/?q=BookType:xxxx”用于检索不同类型的CADAL资源，“http://IP或者域名/cadal/cbook/?q=CreateDate:xxxx”用于检索创建日期包含检索词的CADAL资源等。

2.2.2 组合检索接口及参数

在简单检索接口基础上利用组合检索参数，满足对不同检索子串的CADAL资源记录，可实现针对CADAL资源元数据组合检索调用请求。组合检索接口基本形式为“http://IP或者域名/cadal/cbook/?q=检索子串1 组合检索参数 检索子串2（组合检索参数 检索子串N）”。其中，“检索子串=检索项:检索词”。可用的组合检索参数有AND、OR、fl、score、start、rows、sort、wt，其中score表示返回检索结果的相关度得分，分值没有范围，仅针

对不同的检索条件，具有相对值意义；rows表示定义1次返回多少条记录，默认为10，出于数据安全考虑，每次请求最多返回10条记录；sort表示CADAL接口服务默认按照相关度（score）降序排列检索返回结果，根据需要，可以在调用API的请求中增加参数sort，定义返回结果的显示顺序[5]。如“http://IP或者域名/cadal/cbook/?q=Publisher:人民出版社 AND BookType:minguo”。

2.2.3 接口返回结果

以简单检索为例，使用题名为检索条件，调用接口“http://IP或者域名/cadal/cbook/?q=Title:永嘉县志”，系统默认返回XML格式的检索结果。如返回结果<result name=“response” numFound=“718” start=“0”>中，numFound=“718”表示查询到的相关结果条目数量为718条，start=“0”表示按照相关度得分排序后，返回结果从第1条开始（系统计数从0开始）；标签<doc>与</doc>间的部分为每条资源的具体元数据信息，包括BookNo、BookType等信息。调用程序按需要对接口返回结果进行元数据字段提取、处理并组合出有效命中记录的全文访问链接，最后将相关元数据信息整合到OPAC页面，呈现给读者参考和使用。

2.3 整合检索流程设计

清华大学图书馆OPAC系统与CADAL数字资源平台是两个异构的资源管理系统，借助CADAL元数据开放接口可实现OPAC系统与CADAL数字资源平台元数据交互。其利用开放接口为读者查询所需CADAL数字资源，将检索结果整合到OPAC页面，为读者呈现CADAL数字资源全文访问链接。

读者通过OPAC系统检索资源并获取CADAL相关数字资源包括五个步骤，整合检索处理流程见图1。

（1）读者在清华大学图书馆OPAC系统执行查询操作；（2）通过嵌入OPAC系统的接口调用处理程序，提取读者输入的“题名”“作者”“ISBN”或“关键词”等检索字段和检索值；（3）嵌入OPAC系统中的接口调用处理程序触发接口调用，生成对CADAL元数据的HTTP请求并发送到CADAL开放接口系统；（4）CADAL开放接口系统对请求进行权限、语法等方面的合规性检测后，将检索结果提供给OPAC系统进行后续处理与呈现；（5）OPAC系统接到接口系统的返回结果，对返回数据进行甄别并将有效结果与OPAC系统馆藏检索结果进行资源整合，供读者参考和使用。

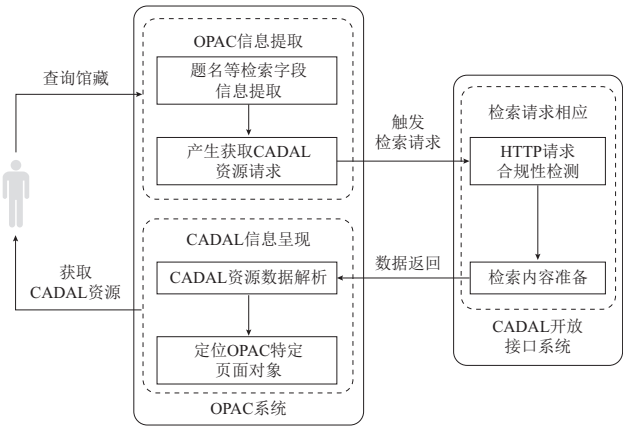

图 1 整合检索处理流程

按照功能划分，整合检索处理流程涉及整合检索逻辑与应用场景部分、接口调用处理程序部分、OPAC页面整合呈现部分；按照处理步骤可细化为四步，即读者检索信息提取与甄别、按照接口规范生成CADAL元数据接口调用请求并发送给接口系统、接收接口返回数据并处理、生成CADAL全文对象访问链接整合到OPAC页面呈现给读者。

3 OPAC系统整合检索CADAL数字资源功能实现

OPAC系统整合检索CADAL数字资源功能实现中的关键细节问题（包括读者在OPAC系统提交检索选项和检索词处理、整合检索逻辑与应用场景设计、接口返回数据处理与分析、CADAL资源全文链接定位、接口调用程序编写中需注意的问题等），以供图书馆同行参考。

3.1 OPAC检索信息处理

根据读者在OPAC系统的检索选项和检索词，可设计不同的整合检索逻辑及应用场景。读者可选择的检索选项包括题名、作者和ISBN等。检索选项为ISBN，则对应的检索逻辑为精确匹配，应用场景为查找CADAL数字图书馆中是否存在ISBN相同的数字资源；检索选项为题名或关键词，对应的检索逻辑为模糊匹配，应用场景为查找CADAL数字图书馆是否存在与检索词相关的数字资源；整合呈现页面可在OPAC搜索结果页面或资源详情页面实现，两种情况对应的接口调用参

数有所差异。搜索结果页面呈现与读者检索词相关的CADAL数字资源，资源详情页面呈现与具体资源相关的CADAL数字资源。整合检索的主要目的是方便读者获取所需资源，尽可能将读者检索的相关资源精确定位并呈现。若读者精确查询，可利用ISBN在CADAL元数据中进行精确定位；若读者通过题名或关键词等字段进行资源查询，只能利用检索词模糊检索CADAL元数据，获得与其查询目标相关的数字资源。在整合检索设计与应用中，可考虑将两种情况综合应用；在OPAC检索结果页面整合呈现时，完全按照读者输入的检索选项和检索值进行接口调用和元数据匹配；在OPAC资源详情页面整合呈现时，考虑CADAL数字资源并非都包含ISBN值，在接口调用时可通过题名字段模糊匹配相关CADAL资源。各图书馆在实际应用中需酌情考虑。

3.2 接口返回数据量

CADAL接口系统基于元数据安全考虑，每次接口调用请求最多返回10条记录。CADAL元数据检索结果按照相关度分值进行排序，检索命中结果记录数大于或等于10条时，默认返回前10条记录；如需获取前10条记录以外的元数据，可通过组合检索参数start和rows指定返回结果，基本格式为“http://IP或域名/cadal/cbook/?q=检索项:检索词&start=*&rows=*”。如调用接口“http://IP或域名/cadal/cbook/?q=BookType:minguo&start=100&rows=5”，则返回检索结果对101—105条元数据排序。具体应用中，受限于OPAC系统页面内容数量、布局和美观考虑以及模糊检索匹配精确度等，通常在目标页面整合记录条目最多为10条。

3.3 接口返回数据选取

CADAL元数据检索结果相关度主要受CADAL元数据质量、接口服务系统分词库以及算法等多重因素共同影响，为保证读者有良好的检索体验，避免出现接口系统对于个别检索词返回相关度低的结果，考虑对接口返回数据增加过滤选取工作。按照精确匹配和模糊匹配要求，将过滤选取工作分为两种情况进行处理：（1）若接口调用时按照ISBN对元数据精确匹配，可直接呈现命中结果，无需过滤操作；（2）若接口调用时无法按照ISBN对元数据精确匹配，需通过读者输入检索词进行模糊匹配，可按照返回结果中相关度分值对命中元数据条目进行过滤操作，选取规则为检索命中记录的相关度分值大于1且高于最大相关度分值的1/2，将满足该条件的结果记录按照相关度得分排序呈现。接口返回数据选取流程见图2，其检索结果整合遵循的是混搭理念。混搭作为Web 2.0的典型应用，指将不同来源的数据和功能无缝组合，形成全新、集成式的服务。清华大学图书馆于2008年开始尝试将混搭理念引入OPAC系统，先后实现在OPAC页面上汇集书封、短信、馆藏地图和多媒体资源等服务，目的是给读者提供多样化信息资源和独特的应用体验。本应用中整合CADAL资源采用的是同样的思路，实现方式是在页面不同分区中呈现不同来源的资源。OPAC资源与CADAL相关资源未合并在页面同一个分区中，在页面底部单独呈现检索命中的CADAL相关资源[6-8]。

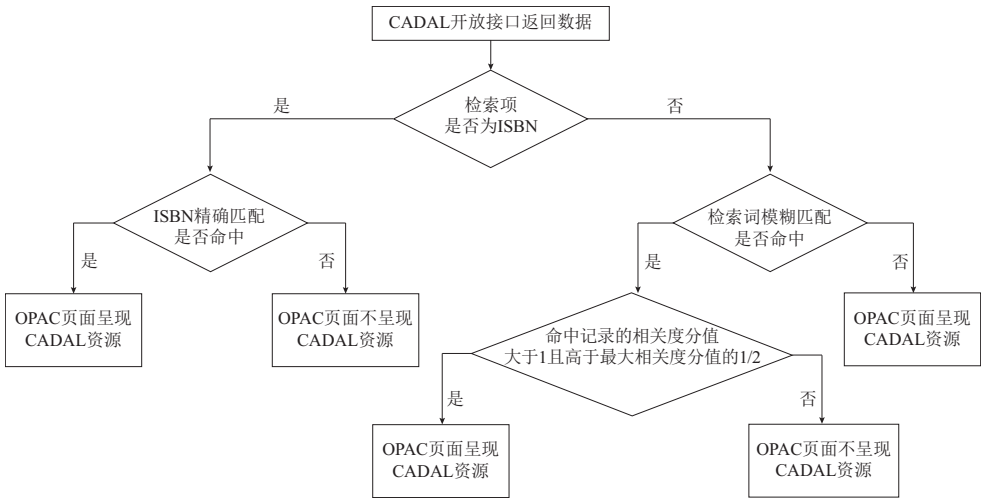

图 2 接口返回数据选取流程

3.4 全文对象链接定位

接口系统返回数据格式默认为XML，也可通过组合检索参数指定其他常用返回格式，如json、python、ruby、php、phps或custom等格式。整合CADAL资源到OPAC页面需提供全文访问链接，以方便读者使用。对于命中记录，提取元数据中的BookNo字段值，以定位CADAL资源的全文对象。所有CADAL资源都具有资源唯一标识BookNo，CADAL数字资源平台中全文对象的URL由URL前缀加上BookNo字段组成，通过解析CADAL元数据，找到CADAL资源的BookNo字段，可组合出相应资源的全文访问链接地址，具体格式为“http://www.cadal.zju.edu.cn/book/”+BookNo。

3.5 接口调用量分析

截至2016年年底，在CADAL管理中心登记试用开放接口系统的成员馆已有30家。开放接口系统采用双服务器负载均衡，当访问量过多时，两台服务器共同分担访问流量，CADAL开放接口服务系统并发量大于500个用户。在高并发用户测试时，事务响应时间最小值为0.343秒，最大值为16.349秒，中间值为2.608秒。实际应用环境中，并发用户量不高，检索结果响应时间均在1秒内，目前接口系统性能和吞吐量可以满足已登记试用成员馆的接口调用需求。2016年，清华大学图书馆OPAC系统共发出264 841次关于民国图书资源的CADAL元数据接口调用请求，日均调用量约725次，平均调用量约为30次/小时。

3.6 接口调用程序开发注意事项

在接口调用程序开发过程的注意事项包括接口调用身份认证、特殊字符处理、URL编码等。

（1）认证。使用CADAL开放接口系统需经过IP和用户白名单双重认证。使用接口系统前须注册接口调用机器的IP地址，当CADAL接口服务器收到HTTP请求时，要先判断该请求是否来自授权IP地址，若IP地址未经授权，则接口调用请求不会被系统处理。另外，基于服务器端编程语言调用接口时，除提供服务器IP地址外，还要向CADAL管理中心申请用户账号；基于客户端开发语言调用接口时，考虑到JavaScript等客户端脚本语言的源代码可见，用户名和密码信息不安全，使用一组仅需IP认证的接口。由于清华大学图书馆OPAC系统存在封闭性，在应用中使用客户端开发语言JavaScript进行接口调用。

（2）特殊字符。调用CADAL提供的开放接口，需遵循相应规则生成合规的HTTP请求。在生成HTTP请求前，需对检索词中的特殊字符进行处理，再提交检索请求。如检索词为“钕玻璃中Na~（3+）离子光吸收温度依从性及钕玻璃光纤温度”，经处理得到“钕玻璃中Na\~\（3\+\）离子光吸收温度依从性及钕玻璃光纤温度”。英文文献的题名、作者等字段经常出现空格，若不预先处理就直接提交检索，CADAL接口系统会误认为是组合检索请求，因此提交检索请求前，需将检索词中空格替换成“AND 检索项:”。

（3）URL编码。URL编码格式采用ASCII码，不能在URL中包含任何非ASCII字符（如中文）。因此，须对调用接口的URL进行转换，生成有效的ASCII字符格式。如“大学”转换后为“%E5%A4%A7%E5%AD%A6”。各种编程语言都有相应的类和方法实现编码功能。不同的浏览器对包含中文的URL处理具有不同的表现，有的浏览器地址栏在显示URL时会自动进行解码（使用UTF-8字符集）。如在Chrome浏览器地址栏中输入“http://IP或者域名/cadal/cbook/?q=Title:大学”，其中的中文可直接显示，但实际发送给服务端的原始URL是经过编码的。

（4）繁简体汉字。CADAL元数据接口系统会对繁简体汉字进行转换处理，在OPAC系统中调用接口部分程序不必考虑繁简体汉字转换工作。

3.7 OPAC系统整合检索CADAL数字资源建设成效

在清华大学图书馆OPAC系统整合检索CADAL数字资源实际应用中，使用JavaScript客户端语言调用CADAL元数据开放接口。程序分析读者检索OPAC的命中记录，“出版发行”字段有“民国”“民國”字样，或者出版时间在1911—1949年的图书会触发调用CADAL元数据开放接口脚本程序，接口调用处理程序将对检索命中返回结果进行处理并整合呈现相关CADAL数字资源。如在清华大学图书馆OPAC系统中基于关键词字段搜索“古文观止”，在结果列表点击“考正古文观止”打开该资源详情页面，获取该资源在清华大学图书馆的馆藏位置信息与数字资源全文链

接，读者点击题名后可在浏览器直接打开全文对象，阅读该资源电子版全文。

清华大学图书馆OPAC系统整合检索CADAL数字资源应用得到广大读者好评，扩充了读者获取资源的渠道，部分民国图书由于馆藏副本较少，给读者借阅造成不便。借助整合检索功能的应用，使读者在馆藏详情页面获取整合呈现的CADAL全文资源，方便读者获取CADAL数字资源平台中民国图书全文电子版，一定程度上缓解了读者借阅问题。

4 结语

设计与开发清华大学图书馆OPAC系统整合检索CADAL数字资源功能，将清华大学图书馆馆藏资源与CADAL丰富的数字资源有效、无缝地关联和整合，为读者提供包含异构资源的检索结果，建立跨资源、跨系统的资源共享环境，更好地为读者服务。该功能不仅提高了CADAL数字资源在清华大学图书馆OPAC系统的利用率，还扩充了读者获取CADAL数字资源的渠道和方式。

各高校图书馆OPAC系统的运行环境、程序开发语言及页面结构等情况各有不同，但CADAL元数据开放接口系统提供了独立、通用、标准化的应用程序接口。其他高校图书馆可通过标准化接口调用与程序处理，将CADAL相关数字资源元数据信息整合到自身OPAC系统。针对现有OPAC系统整合检索CADAL数字资源的读者使用体验来看，仍存在不足，需后续改进。CADAL数字资源平台全文访问需要用户登陆认证，图书馆读者发现感兴趣的资源后，需输入CADAL数字资源平台的用户名和密码，在认证成功后才可查看全文。日后应避免读者在不同系统间多次登陆的问题，优化读者使用体验，后续可考虑通过CADAL成员图书馆IP地址免认证或实现OPAC系统与CADAL数字资源平台统一认证功能来解决上述问题。

参考文献

[1] 周虹,张蓓,窦天芳,等.清华大学图书馆OPAC书封服务的设计与实现[J].现代图书情报技术,2008(8):84-87.

[2] CADAL.资源服务动态[EB/OL].[2017-05-24].http://www.CADAL.cn/.

[3] CADAL.CADAL数字资源入库量、在线量分类统计[EB/OL].[2017-05-24].http://www.CADAL.cn/zydt/index1512.htm.

[4] 李书宁,王琼.图书馆资源发现应用OpenAPI标准化研究[J].图书情报工作,2012,56(7):16-20.

[5] SMILEY D,PUGH E.Apache Solr 3 enterprise searchserver[M].Birmingham:Packt Publishing Ltd,2011.

[6] 窦天芳,姜爱蓉,陈武.以Exlibris & Metalib为例谈整合检索的几个关键技术及应用[J].情报科学,2007,25(8):1235-1239.

[7] 翟晓娟,聂娜.满足用户个体需求的图书馆开放平台设计——基于OpenAPI、App、Mashup、SOA的集成实践应用[J].大学图书馆学报,2011(6):26-32.

[8] 周朝阳,王时绘.面向服务的资源整合检索系统研究与实现[J].现代情报,2009,29(9):175-178.

作者简介

远红亮，男，1982年生，硕士，馆员，研究方向：数字图书馆、图书馆IT及信息化建设，E-mail：yuanhl@lib.tsinghua.edu.cn。
张蓓，女，1979年生，硕士，副研究馆员，研究方向：数字图书馆、图书馆IT及信息化建设。
张成昱，男，1966年生，博士，副研究馆员，研究方向：数字图书馆、图书馆IT及信息化建设。
周虹，女，1976年生，硕士，副研究馆员，研究方向：数字图书馆、图书馆IT及信息化建设。

Research on CADAL Digital Resource Integration Retrieval: Take Tsinghua University Library OPAC System for Example

YUAN HongLiang, ZHANG Bei, ZHANG ChengYu, ZHOU Hong
(Tsinghua University Library, Beijing 100084, China)

Abstract: Through calling CADAL metadata open interfaces system, Tsinghua University Library OPAC system can search and integrate metadata of CADAL digital resources, and publish record full text access links in the result page, which could enhance readers experiences and improve CADAL digital resources rendering and utilization, and readers could retrieve and utilize related CADAL digital resources conveniently when they are using OPAC system. This paper focuses on design and implementation of CADAC metadata search and integration in OPAC system, which could be a good reference for the same type of library or system designers.

Keywords: OPAC; CADAL; Metadata; OpenAPI

（收稿日期：2017-06-26）