基于开源软件实现馆藏数字资源整合与统一检索*

陈 和 王 爽
(厦门大学图书馆 厦门 361005)

【摘要】厦门大学图书馆采用开源软件 PKP Harvester2对不同的馆藏数字资源进行基于 OAI-PMH 协议的元数据收割,实现馆藏多种数字资源的整合,同时提供一站式的统一检索服务。主要介绍 PKP Harvester2系统的功能特点、实施过程以及系统本地化过程。

【关键词】数字资源整合 OAI-PMH PKP Harvester2 系统本地化 厦门大学
【分类号】G250.76

Implementation of Digital Resource Integration and a Unified Search Based on Open Source Software

Chen He Wang Shuang
(Xiamen University Library, Xiamen 361005, China)

【Abstract】Xiamen university library harvests metadata from different digital resources based on the OAI-PMH with the open source software PKP harvester2, and achieves a collection of the integration of various digital resources, as well as providing a unified one-step search service. The paper focuses on the features, implementation process and the localization process of the PKP harvester2.

【Keywords】Digital resource integration OAI-PMH PKP harvester2 System localization Xiamen university

1 引 言

随着数字图书馆建设的推进,数字化馆藏在整个馆藏中的比例不断增大,自建特色数字资源和购买的各种学术数字资源也不断增多。馆藏数字资源的增加,一方面为教学科研人员及学生用户提供了丰富的文献资源,保障他们能够掌握全面和前沿的专业学术资源;另一方面,教学科研人员及学生用户为获得比较全面的文献资源,不得不在这些资源或数据库中分别检索。由于这些数字资源建设的不同步以及采用技术的不同,各种数字资源都有各自的数据结构、组织方式、查询方式以及显示界面,为此需要花费大量的时间和精力去掌握这些数据库的检索方法以及在不同的数据库中重复提出相同的检索请求。厦门大学图书馆本着高效、快捷、简约的原则,在运用开源软件 PKP Open Archives Harvester2实现馆藏数字资源整合并提供统一检索方面作了一次实践探索。

2 PKP Open Archives Harvester2系统简介

PKP Open Archives Harvester2(简称 PKP Harvester2)是一款由加拿大公共知识项目组(Public Knowledge Project

收稿日期:2009-04-08
收修改稿日期:2009-05-14
* 本文系网络时代的科技论文快速共享研究基金项目"从'中国科技论文在线'到高校 IR 联盟之服务创新模式研究"(项目编号:2008104)的研究成果之一。

PKP)开发, 基于 OAI- PMH协议, 集成了收割元数据、索引元数据、存储元数据和提供统一检索的开源软件。PKP项目受到致力于扩大和提高开放获取研究的联邦基金资助[1]。PKP Harvester2是该项目组基于旧版 PKP Harvester作了较大的修改和改进而成, 寓意第二版。

2.1 软件特点[1]

(1)能够收割基于 OAI接口的多种不同格式(Schema)的元数据, 包括未受限的 Dublin Core、PKP Dublin Core扩展、MODS和 MARCXML。这些格式在系统中都是以插件的方式进行支持。

(2)灵活的检索界面, 包括简单检索和高级检索。

(3)通过结合使用专题(SetSpec)和时间戳(Timestamps)可以执行颗粒化的数据收割。

(4)元数据在收割后索引前可以进行过滤或标准化操作。

(5)用户界面采用 CSS和基于模板的 HTML语言设计, 可以根据需求进行定制。

(6)高度可扩展检索, 建立了用于检索的索引和反向索引。

(7)可以通过增加插件的形式增加收割协议和元数据格式。

(8)内容敏感在线支持。

(9)灵活的收割设置, 包括 Web界面收割和文本界面收割。

2.2 资源整合原理

如图 1所示, PKP Harvester2通过收割基于 OAI-PMH协议的异构数字资源(资源 1、资源 2、资源 3、资源 n)的元数据, 实现不同数字资源的元数据集中存储, 即构建成“元数据仓储”。在此基础上, 对元数据建立索引和反向索引, 提供资源浏览、检索和其他功能, 给用户一个新的、统一的应用检索平台。用户通过此平台可以一站式地检索各种异构数字资源, 最终达到资源整合与统一检索的目的。

3 系统实施过程

3.1 数字资源准备

从 PKP Harvester2系统的资源整合原理可以知道, 被整合的数字资源必须提供 OAI接口才能被整合。厦门大学图书馆自建数字资源“厦门大学博硕士学位论文库”和“厦门大学学术典藏库”目前已经提供了 OAI接口, 采用的元数据方案为 Dublin Core元数据扩展。对于其他未提供 OAI接口的数字资源参照现有的 OAI接口通过 ASP编程实现, 并且采用 Dublin Core元数据扩展方案。

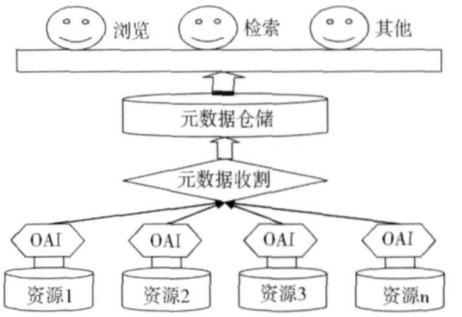

图 1　PKP Harvester2资源整合原理

3.2 PKP Harvester2的安装

只要系统环境配置妥当, PKP Harvester2的安装相对比较简单。需要注意的是, 安装文件解压缩后, config inc php文件和 publis、cache两个文件夹及子文件夹需要添加所有用户的“写”权限(针对 Linux操作系统)。MySQL数据库采用的字符集、客户端字符集和连接字符集均设置为“UTF- 8”格式。

3.3 PKP Harvester2的管理

(1)系统属性管理

PKP Harvester2既是资源整合与统一检索平台, 也是对外服务的一个网站, 因此, 需要对网站的相关属性进行管理。PKP Harvester2软件提供了如下网站属性: 网站名称、网站简介、网站相关描述、自定义的网站 Logo图片、网站 CSS文件、网站语种、功能插件以及管理员相关信息等。

(2)数据源管理

[1] 数据源添加

数据源(Archive)在本系统中是指提供了 OAI接口的将要被整合的馆藏数字资源。在添加数据源时需要提供数据源的名称、相关描述、访问数据源的 URL、开放 ID、接口类型、OAI接口的 URL、管理员 Email、索引方式、元数据格式等。

④元数据收割

添加完数据源后, 选择需要收割的专题(默认是全部专题)和收割的起止时间(精确到日期), 系统将自动把指定时间段内的专题元数据收割并保存到系统的元数据仓储数据库中, 同时完成元数据的索引供检索之用。

3.4 数字资源浏览与检索

待完成数字资源的元数据收割后,系统即可提供被整合资源的浏览和检索服务。在资源浏览界面,可以查看已经被整合的数据源列表以及每个数据源拥有的元数据条目数。点击相应的数据源名称可以按题名或日期浏览该数据源的所有条目的元数据。另外,用户可以通过系统提供的简单检索或高级检索方式,对系统进行统一检索。

4 PKP Harvester2的本地化

PKP Harvester2实施后基本上实现了馆藏数字资源的元数据整合,并提供了资源浏览与检索服务,可以满足普通用户的文献检索需求,其系统界面如图 2所示:

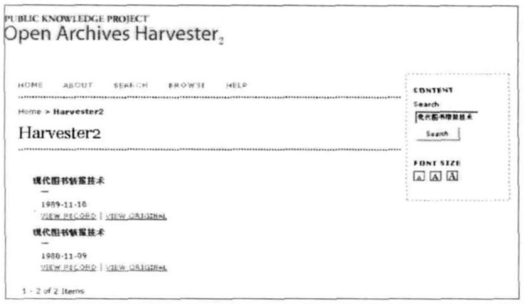

图 2 本地化之前的 PKP Harvester2

为了更好地提升系统的易用性,本案例对 PKP Harvester2进行本地化方面的改进,系统本地化后界面如图 3所示:

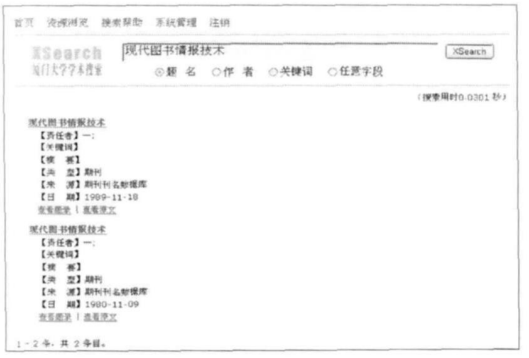

图 3 本地化之后的 PKP Harvester2

4.1 系统本地化的思路

由于 Google操作的易用性、人性化和简约化,以及资源方面的优势,成为多数人所钟爱的搜索引擎,其操作模式已经深入人心并成为了一种操作习惯。因此系统本地化将尽量与 Google风格相靠拢,尽量做到系统的人性化和简约化,提升人机交互的友好性,节约用户学习和使用系统的时间。

如图 4所示,用户与系统交互的过程主要包括:用户从检索入口提交检索需求,系统对检索参数进行规范后进入元数据仓储检索,检索结果经过一定处理后反馈给用户。本案例主要对检索入口、参数处理、结果处理等环节进行了相应的修改和改进。

图 4 系统人机交互过程

4.2 系统界面汉化

系统默认的界面语言为英文,为提高系统的人性化,适应国内本地用户的使用习惯,本案例对系统的界面语言进行了汉化处理。在 PKP Harvester2系统中,系统的界面语言文件与程序内容文件是独立分开设计的,系统汉化只需要对下列文件进行汉化即可[2]。

[harvester] /locale/en_US/locale.xml
[harvester] /dbscripts/xml/data/locale/en_US/email_templates_data.xml
[harvester] /help/en_US/
[harvester] /rt/en_US/

若需要支持英文和中文双语界面,汉化后的文件需要放在上述相关目录下新建立的 zh_CN目录中,同时在[harvester] /registry/locales.xml文件中添加“<locales><locale key="zh_CN" name="中文"/></locales>”语言注册信息,最后在网站管理界面中安装中文语种。

4.3 检索功能和检索方式改进

系统默认提供了两种 Web检索方式,即高级检索和简单检索,如图 5所示。高级检索提供了多组输入框,可以指定数据源和选择多个检索元数据项进行组合检索,如图 5左侧栏所示。简单检索只提供一个输入框,为全字段检索,即检索时搜索引擎将检索词与所有元数据项进行匹配,如图 5右侧栏所示。

高级检索的检索结果通常正是用户所需要的结果,但是检索操作相对比较繁琐。简单检索操作只要一次输入检索词即可进行检索,但是检索结果数量往往较多,而且很多检索结果是不相关结果,对用户甄别

有用信息造成了干扰。为了在简单检索与高级检索之间找到平衡点, 本案例根据 PKP Harvester2提供的功能设计了基于 URL的检索方式。

图 5　PKP Harvester2的高级检索与简单检索界面

(1)相关语法

PKP Harvester2系统为了方便软件开发人员从其他 Web应用程序动态产生链接到 PKP Harvester2的内容, 提供了一个简单的类似 OpenURL标准的用于执行检索的 GET URL语法, 该 URL的语法格式如下[3]:

http://domain_name/harvester/index.php/search/byUrl

若要指定执行一个检索, 只要在上述 URL后加上相应的参数即可。例如:

http://domain_name/harvester/index.php/search/byUrl?archive=test&title=name

上述 URL语法表示检索数据源开放 ID为“test”, 并且 title字段值为“name”的所有条目。

(2)检索入口的设计

检索入口是用户向系统提交检索方式和检索词的入口。针对 PKP Harvester2提供的功能, 本案例在保留原来系统的简单检索功能的基础上, 添加了按题名、按作者、按关键词三种检索方式。这种设计方式一方面给用户提供了多种检索途径, 另一方面检索内容更有针对性, 提高了查准率。

检索入口设计语言为 HTML, 包括三个部分: 输入框、检索方式选择和提交按钮。检索方式选择有两种设计方式, 分别是下拉菜单选择方式和单选按钮组方式。前者设计可以为界面节约空间, 但是需要执行两个操作步骤; 后者占用界面比较多的空间, 但是选项直观, 只需要执行 1个操作步骤。根据本地化原则, 采用后者设计方式, 设计效果如图 3所示。

4.4　检索参数处理

检索入口提交的参数包括检索方式和检索词, 根据 PKP Harvester2的 GET URL语法要求, 需要转换为检索标准 URL中的相关参数, 这部分处理是由 transform.php文件完成的, PKP Harvester2系统中并没有处理此过程的文件, 需要另外设计。transform.php文件的主要代码如下:

```
switch ($field)
{
case 'any': /* 任意字段检索* /
 $url='http://localhost/harvester/index.php/search/results?query='.$searcharea'';
break;
case 'title': /* 按题名字段检索* /
$url='http://localhost/harvester/index.php/search/byUrl?title='.$searcharea'';
break;
case 'creator': /* 按作者字段检索* /
$url='http://localhost/harvester/index.php/search/byUrl?creator='.$searcharea'';
break;
case 'subject': /* 按关键词字段检索* /
 $url='http://localhost/harvester/index.php/search/byUrl?subject='.$searcharea'';
break;
}
```

4.5　条目显示优化

条目显示优化属于结果处理部分。系统默认的条目显示效果如图 2所示, 结果包括条目的三个元数据项和两个查看链接, 分别是题名(Title)、责任者(Creator)、日期(Date), 以及查看题录(View Record)和查看原文(View Original)。对于用户而言, 这样的显示信息过于简单, 不容易判断该条目的“有用性”或价值, 特别是在有多个相似条目一起显示时, 更难取舍。如果需要了解比较详细的条目信息, 用户只能点击“查看题录”或直接点击“查看原文”来进行查阅, 从易用性角度来看, 此操作增加了用户的操作和使用时间。

本案例对系统的条目显示进行了优化, 条目显示结果包括条目的多个元数据项, 即题名、责任者、关键词、摘要、类型、来源和日期。另外, 为方便用户查看原文, 对题名部分还增加了查看原文链接。涉及条目显示优化的主要文件是[harvester]/plugins/schemas/dc/summary.tpl, 优化后的显示效果如图 3所示。

4.6　元数据收割方式改进

PKP Harvester2系统默认的元数据收割方式是在

数据源管理的 Web 界面中, 输入需要收割的起止时间, 然后再进行收割。此元数据收割方式存在以下不足:

(1)收割时间精确度不够。默认的只能精确到日期, 不能精确到时、分、秒。对于时间戳精确到时、分、秒, 并且需要按小于日期的时间间隔来收割的数据源来说, Web 界面收割就不能完成此收割任务。

(2)不能显示收割过程。默认的收割过程是在系统后台进行的, 用户不能查看收割进度状态以及收割过程中出现的错误或警告信息。当数据源的数据量巨大时, 往往会发生时间过长而中断的现象, Web 界面收割无法查看和处理此故障。

(3)不能实现自动定时收割。默认的 Web 界面收割需要管理员每次登录管理界面, 然后输入收割的起止时间进行收割。对于静态的数据源 (即数据量不再增加或减少的数据源)来说, 这种收割方式不存在问题, 但是对于类似馆藏书目数据源来说, 其数据量每天都在增加, 这种情况需要管理员每天手工操作进行元数据收割, 这样的重复操作增加了管理员的工作量。

针对上述 Web界面收割方式的不足, PKP Harvester2系统还设计了在命令模式下运行的收割工具, 其用法如下[3]:

php[harvester] /tools/harvest php[archive ID] [from = YYYY - MM - DD] [until= YYYY - MM - DD] [set= setSpec] [flags]

archive ID是指数据源在数据库中的 ID号, 但是有两个特例, 如果其值为“ list”则将列表显示系统中已有的数据源的 ID号、数据源名称以及目前已收割的条目数; 如果其值为“ all”, 则表示将对所有的数据源进行操作。

from = YYYY - MM - DD、 until= YYYY - MM - DD是分别指定收割的起始时间和终止时间, 指定的时间格式可以是精确到日期的“YYYY - MM - DD”格式, 也可以是精确到秒的“ YYYY - MM - DDTHH: MM: SSZ”格式。若“YYYY - MM - DD”被替换为“ now”, 则表示当前时间; 若“ YYYY - MM - DD”被替换为“ last”, 则表示此数据源最后一次执行元数据收割的时间。

set= setSpec是指定只收割数据源的某个专题。

flags有 4个可选值, 分别是 verbose、 flush、 usage、 skipIndexing。 verbose表示将显示收割进程状态信息; flush表示在收割元数据之前将删除原来收割的元数据; usage表示显示指定数据源的额外使用信息; skip-Indexing表示收割元数据时将跳过删除元数据和创建索引过程 (系统默认是先收割 50条元数据, 然后创建索引, 然后再收割, 再建索引)。

例如: php harvest php 1 from = 2009- 03- 25T23: 00: 00Z until= 2009- 04- 01T01: 00: 00Z verbose flush

表示收割 ID号为 1, 起始标准时间为“ 2009- 03- 25T23 00: 00Z”, 终止标准时间为“ 2009- 04- 01T01: 00: 00Z”内的所有元数据, 并显示收割过程中的相关信息, 而此前收割的元数据将会被删除。

在本案例中, 为了尽量减少人工干扰收割过程, 创建了自动收割脚本并置于 Linux的 cron任务中。 cron任务列表如下:

06* * * (cd [harvester] /tools/; php harvest php all from = last until= now verbose > > log_harvester log)

此任务表示系统每天早晨 6点, 对所有的数据源进行元数据收割, 收割的起止时间为上次收割时间到现在的时间, 收割的信息记录到 log_harvester log文件中。

通过自动执行此任务, 就使系统中的元数据每天得到更新, 基本上与数据源的数据保持同步。另外, 管理员可以通过查看 log_harvester log文件来了解各个数据源的元数据收割情况, 对相关的警告或错误信息进行处理。

除了上述的本地化工作外, 本案例还对网站 CSS、 Logo图片等进行了修改和优化。

5 系统的不足之处

诚然, PKP Harvester2在资源整合领域是一个非常优秀的开源软件, 但是也有其不足之处。通过本案例的实践, 发现 PKP Harvester2软件目前主要存在以下不足:

(1)被整合资源有限。从 PKP Harvester2的资源整合原理可以知道, 被整合的数字资源需要提供 OAI接口或者可以通过编程添加。当前图书馆的数字资源种类繁多, 大体上可以分为自建资源和购买的商业资源。自建资源基本上可以通过编程实现 OAI接口。但是商业资源出于商业利益的考虑, 很多资源并没有或不开放其 OAI接口, 或者其底层数据库是封闭的, 无法通过编程实现 OAI接口。而资源实体不在本地, 且只有访问权限的国外商业资源, 就更谈不上 OAI接口了。

(2)中文兼容性不足。 PKP Harvester2系统为了提

高检索速度和效率, 对收割过来的元数据进行了索引和反向索引。在作索引时, 需要对字符进行切分词处理, MySQL先天对中文切分词不支持, 由此决定了 PKP Harvester2对中文兼容性不足。表现在系统功能上, 就是在用中文的自然词进行检索时常常会发生检索不到结果的情形。

目前一个解决办法是采用通配符 (英文半角的“* ”)的方式进行检索, 但是这样的检索方式不能用于提高检索速度的索引, 需要通过扫描全表的方式进行检索, 大大降低了检索速度。可喜的是, 新版的 PKP Harvester2将增加 Lucence作为其检索引擎, 中文兼容性方面可能会得到改善。

(3)检索结果处理有待改进。系统默认对检索出的结果按条目 ID进行排序, 没有对检索结果进行去重或归并处理。

6 结　语

到目前为止, PKP Harvester2系统已经整合了厦门大学书目数据库、厦门大学硕博士学位论文库、厦门大学学术典藏库等 13种馆藏数字资源, 条目数量将近 2 000 000条。网站日均访问量为 500- 600人次。本案例只是在资源整合方面运用开源软件作了一个初步的实践探索, 在整合的内容和形式上还有待进一步充实和完善。

参考文献:

[1] Open Archives Harvester[EB /OL]. [2009- 04- 05]. http //pkp sfu ca/? q= harvester

[2] Simon Fraser University Library. PKP Harvester2 (Version 2 0) Technical Reference [EB /OL] [2009- 04- 05]. http: //pkp sfu ca/PKP Harvester2 /TechnicalReference pdf

[3] Simon Fraser University Library. PKP Harvester2 in an Hour (Version 2 0) Administrator' s Guide [EB /OL]. [2009- 04- 05]. http //pkp sfu ca/PKP Harvester2 /AdminGuide pdf

(作者 E- mail: xmu_chen@ 163 com)

电子档案袋的新视角

2009年 6月, 在伍尔弗汉普顿大学召开了题为“电子档案的故事: 基于电子档案的学习”会议。会议为期两天, JISC的具有电子档案式的出版工作就是该会议的一个成果。

会议的宗旨是从包括普通学习者、专业工作者、相关机构协会成员以及终身学习工作者在内的各种不同角度, 探讨电子档案系统在使用过程中的相关问题。

《学习和教学中的创新》课程的首席讲师, 来自伍尔弗汉普顿大学的 Julie Hughes教授说: “JISC参与了今年为同仁们提供相互学习机会的‘电子档案袋系统式学习’会议, 对此我感到非常高兴”。

JISC电子学习项目主管 Lisa Gray说: “通过应用电子档案袋系统, 主办机构降低损耗开销, 这表明大学之间分享最优计划和行动的确存在优势。如果高等教育机构间的共享合作水平能提高一个百分点, 所产生的价值将超过 1. 32亿英镑。”

JISC代表整个未来教育和高等教育机构实施了电子学习项目, 在如何使用 Web2 0和在线学习工具 (这些工具能为所有人拓展学习、教育和研究)的方面积极向全体人民提供建议和指导。

(编译自: JISC' s e- Portfolio Work Inspiration for Conference [2009- 05- 14]. http //www. jisc ac. uk/news/stories/2009/05 /eportfolios aspx)

(本刊讯)