

合肥市少年儿童图书馆

网事典藏经验交流

第一章：项目建设流程

第一节、采集

- 采集准备：先进行网页的信息采集，网页采集的标准是政府网站为主。
- 市馆提交省馆初审，省馆初审后，连同初审意见一并提交给国家图书馆审核，由国家图书馆出具审核意见。
- 举例如下：

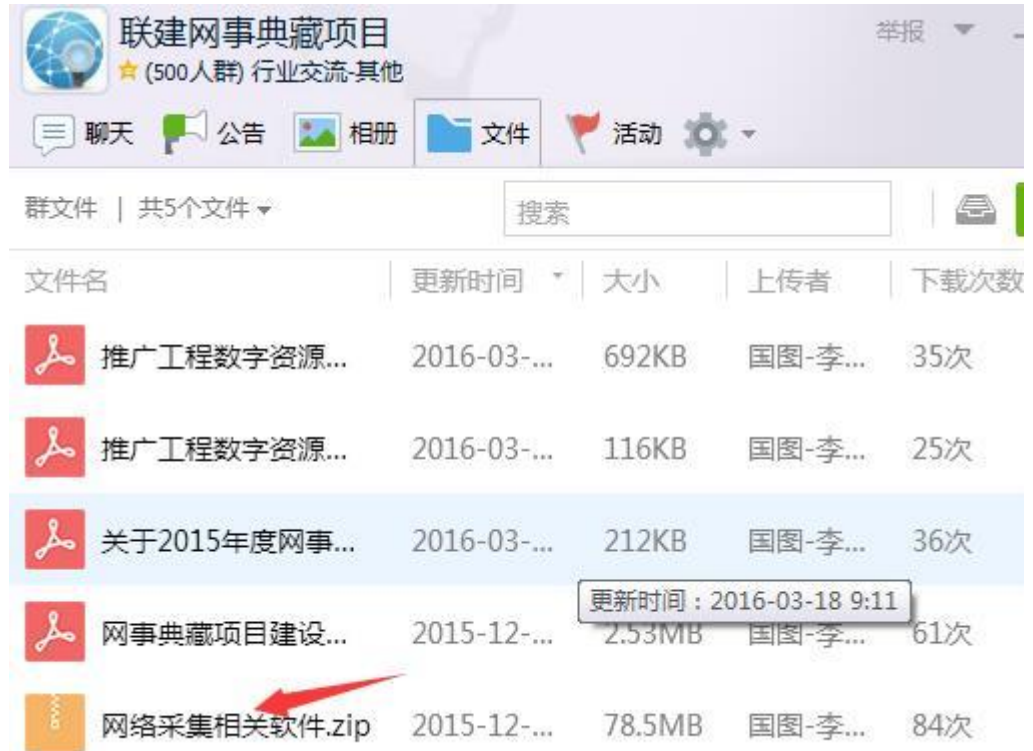
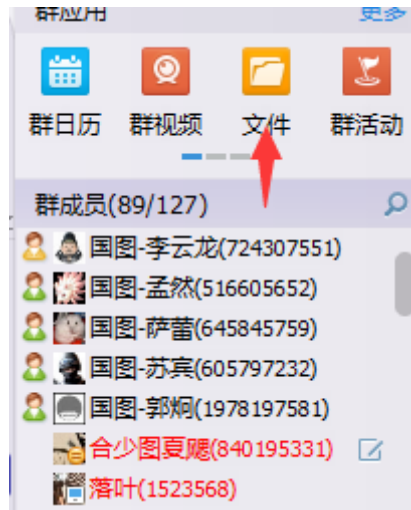
网站名称	网页地址
合肥市人民政府	http://www.hefei.gov.cn/
合肥市委宣传部	http://swxcb.hefei.gov.cn/
合肥市统计局	http://t.j.hefei.gov.cn/
合肥市审计局	http://s.j.hefei.gov.cn/
合肥市规划局	http://www.hfsgjh.gov.cn/
合肥市城市管理局	http://www.hfsr.gov.cn/
合肥市环境保护局	http://www.hfepb.gov.cn/
合肥市监察局	http://www.hfsjw.gov.cn/
合肥文广新局	http://swhj.hefei.gov.cn/
合肥卫生局	http://swsj.hefei.gov.cn/
合肥计生委	http://sisw.hefei.gov.cn/
合肥体育局	http://stvj.hefei.gov.cn/
合肥经开区网站	http://www.hetda.gov.cn/

第一节、采集

- 资源采集：根据采集列表，利用网络采集软件（如 **heritrix**），对政府网站进行全面采集，要求所采集的文件包含采集列表中政府网站的全部内容，但不包括论坛等需要链接后台数据库的内容。所采集的文档格式遵循 **warc1.0** 标准。
- 数据发布：将采集到的文档（**warc**文档）数据进行发布。保证页面内容都能正常打开，且与原网站保持一致。

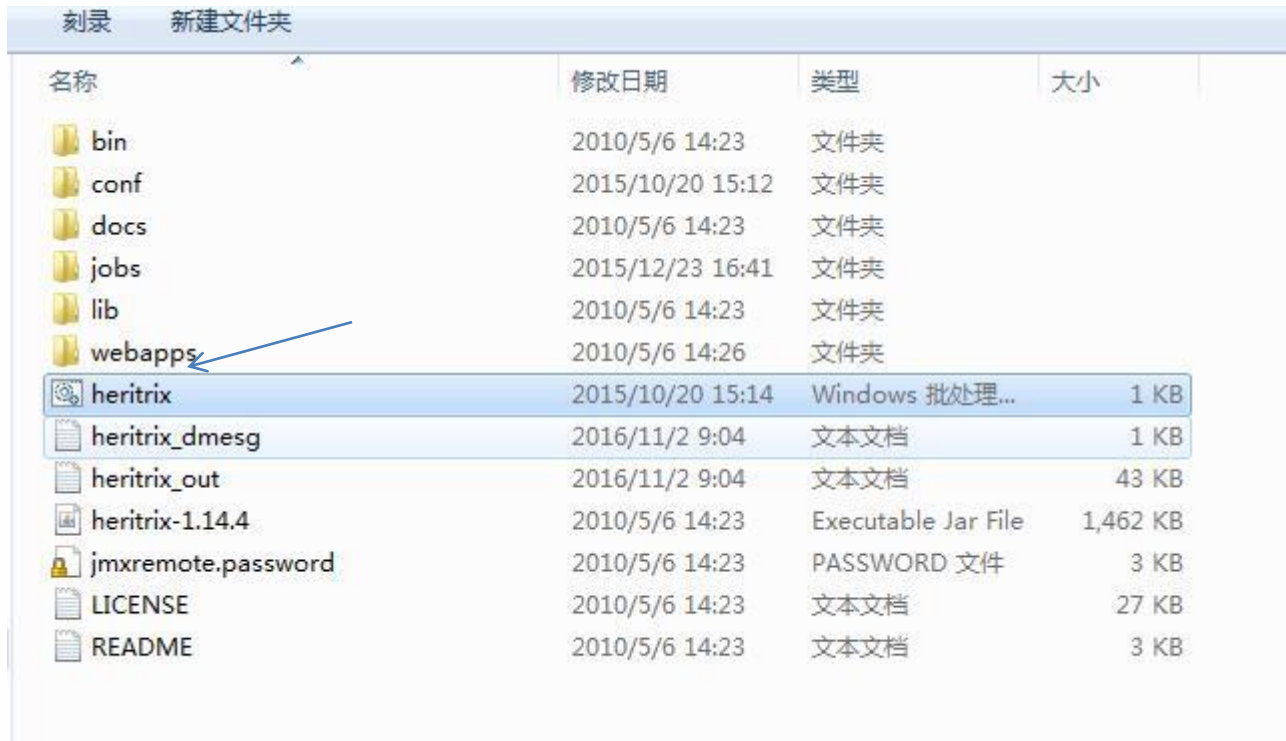
第一节、采集

- 网络采集软件heritrix的使用方法:
- 软件获取办法: 可通过百度搜索或者群共享文件里面下载。
QQ群: 365776635



第一节、采集

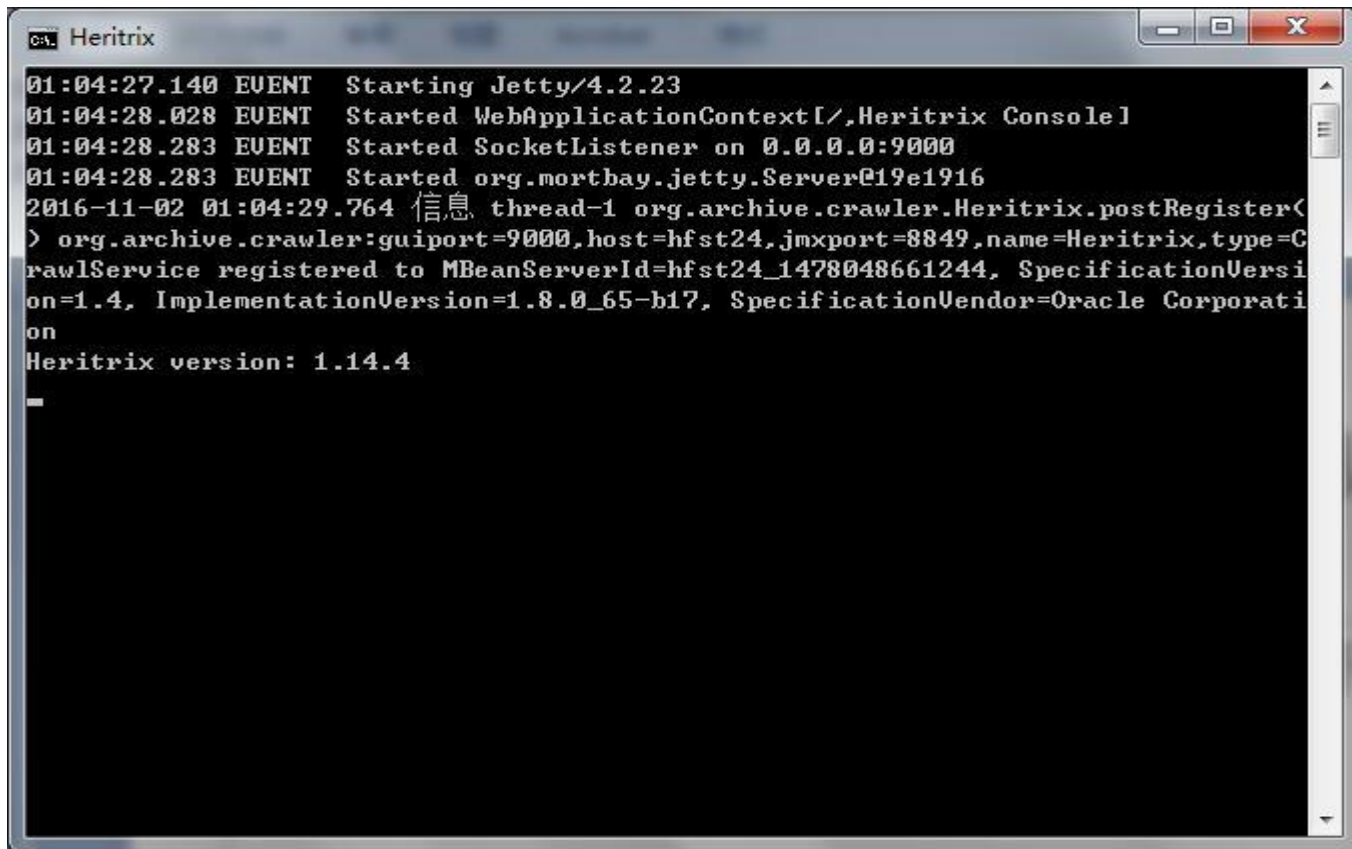
- 安装java，相应的版本可以在网站上下
- 启动heritrix。



名称	修改日期	类型	大小
bin	2010/5/6 14:23	文件夹	
conf	2015/10/20 15:12	文件夹	
docs	2010/5/6 14:23	文件夹	
jobs	2015/12/23 16:41	文件夹	
lib	2010/5/6 14:23	文件夹	
webapps	2010/5/6 14:26	文件夹	
heritrix	2015/10/20 15:14	Windows 批处理...	1 KB
heritrix_dmesg	2016/11/2 9:04	文本文档	1 KB
heritrix_out	2016/11/2 9:04	文本文档	43 KB
heritrix-1.14.4	2010/5/6 14:23	Executable Jar File	1,462 KB
jmxremote.password	2010/5/6 14:23	PASSWORD 文件	3 KB
LICENSE	2010/5/6 14:23	文本文档	27 KB
README	2010/5/6 14:23	文本文档	3 KB

第一节、采集

- 点击heritrix.bat文件。显示的信息如下：



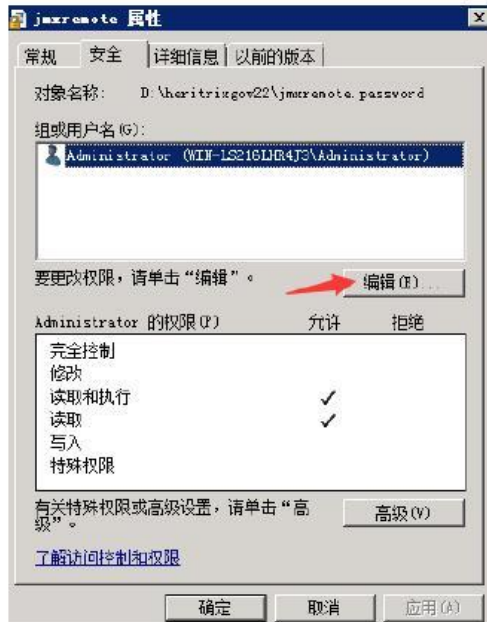
```
ca. Heritrix
01:04:27.140 EVENT Starting Jetty/4.2.23
01:04:28.028 EVENT Started WebApplicationContext[/,Heritrix Console]
01:04:28.283 EVENT Started SocketListener on 0.0.0.0:9000
01:04:28.283 EVENT Started org.mortbay.jetty.Server@19e1916
2016-11-02 01:04:29.764 信息 thread-1 org.archive.crawler.Heritrix.postRegister(
) org.archive.crawler:guiport=9000,host=hfst24,jmxport=8849,name=Heritrix,type=C
rawlService registered to MBeanServerId=hfst24_1478048661244, SpecificationVersi
on=1.4, ImplementationVersion=1.8.0_65-b17, SpecificationVendor=Oracle Corporati
on
Heritrix version: 1.14.4
-
```

第一节、采集

- 如果出现错误，那么就需要删除jmxremote文件的权限。

启动Heritrix (问题)

需要删除jmxremote文件的权限



第一节、采集

- 打开IE浏览器，在地址栏输入<http://127.0.0.1:9000>,输入用户名密码（均为admin），点击login登录。



第一节、采集

- 创建任务实例:

HERITRIX Status as of 十一月. 2, 2016 01:24:56 GMT Alerts: no ale
HOLDING JOBS
Setup 1 jobs pending, 1 completed

Console Jobs Profiles Logs Reports Setup Help

Heritrix Setup

[Local Instance](#)

Choose an instance of Heritrix to manage, or create new instances.

[Web UI Preferences](#)

Change admin password, or change the site's icon.

Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=9000,host=hf.st24

第一节、采集

- 给任务实例起个名，简单的比如1，2，3等。

Instances 1 jobs pending, 1 completed

Console	Jobs	Profiles	Logs	Reports	Setup	Help
---------	------	----------	------	---------	-------	------

Local Heritrix Instances

Use this page to instantiate new instances of Heritrix.

Below is a listing of the Heritrix instances currently running locally. To create a newly created instance, select the instance name in the below list. This sets the UI. This will destroy the instance cleanly terminating any running jobs. Note, you cannot instance to juggle.

Instance Name

[guiport=9000, host=hfst24, jmxport=8849, name=Heritrix, type=CrawlService](#) isRunning=false isCrawling

Name of new Heritrix instance:

第一节、采集

- 选中该实例：

HERITRIX Status as of 十一月. 2, 2016 01:28:50 GMT Alerts: no alerts
HOLDING JOBS
Instances 1 jobs pending, 1 completed

[Console](#) [Jobs](#) [Profiles](#) [Logs](#) [Reports](#) [Setup](#) [Help](#)

Local Heritrix Instances

Use this page to instantiate new instances of Heritrix.

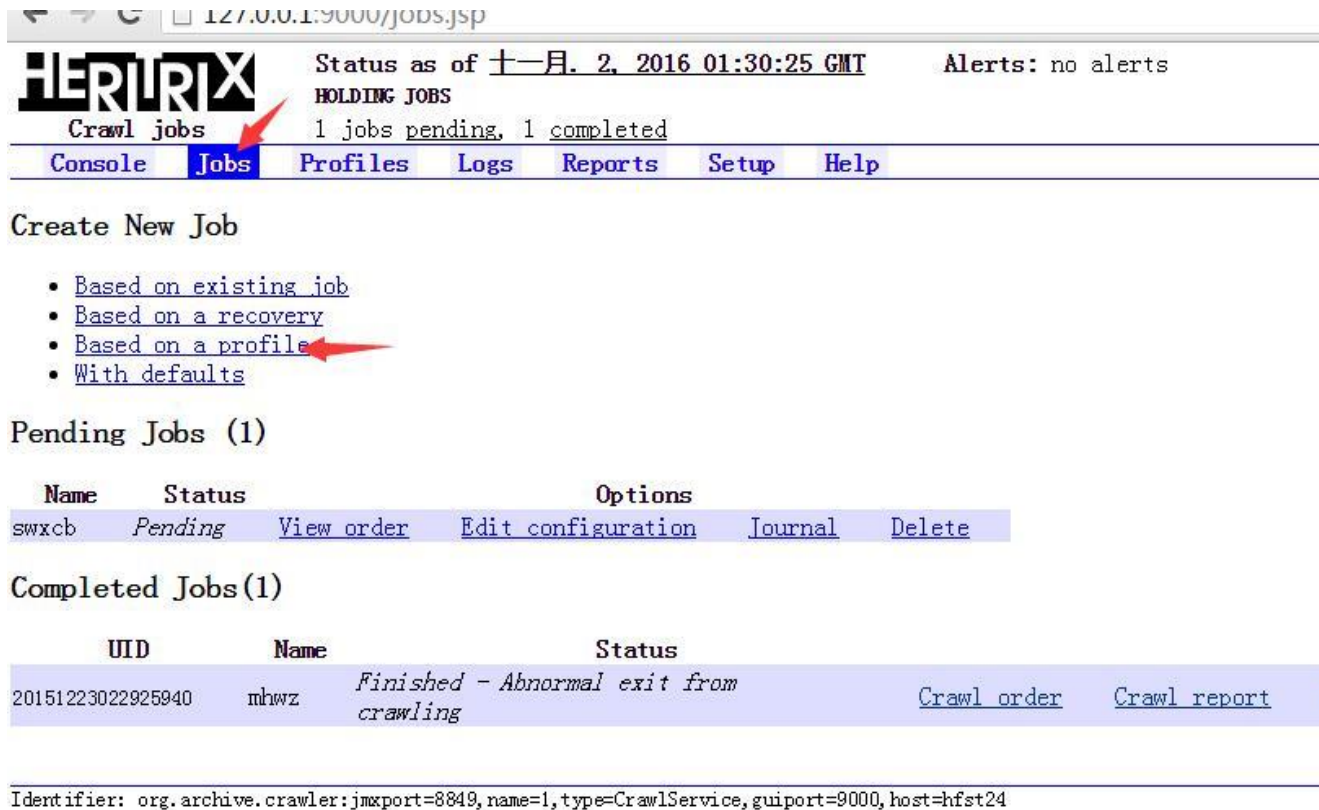
Below is a listing of the Heritrix instances currently running locally. To create a new instance, fill in the textbox below a newly created instance, select the instance name in the below list. This sets the UI running against the selected instance. This will destroy the instance cleanly terminating any running jobs. Note, you cannot delete all Heritrix instances. The UI g instance to juggle.

Instance Name	Status
guiport=9000, host=hfst24, jmxport=8849, name=Heritrix, type=CrawlService	isRunning=false isCrawling=false alertCount=0 newAlertCount=0 <input type="button" value="Delete"/>
guiport=9000, host=hfst24, jmxport=8849, name=1, type=CrawlService	isRunning=false isCrawling=false alertCount=0 newAlertCount=0 <input type="button" value="Delete"/>

Name of new Heritrix instance:

第一节、采集

- 创建采集任务:



The screenshot shows the Heritrix web interface. At the top, the status is "Status as of 十一月. 2, 2016 01:30:25 GMT" and "Alerts: no alerts". Below this, it says "HOLDING JOBS" and "Crawl jobs 1 jobs pending, 1 completed". A navigation menu includes "Console", "Jobs", "Profiles", "Logs", "Reports", "Setup", and "Help".

Under "Create New Job", there are four options:

- [Based on existing job](#)
- [Based on a recovery](#)
- [Based on a profile](#) (indicated by a red arrow)
- [With defaults](#)

Below this, there are two tables:

Pending Jobs (1)

Name	Status	Options			
swxcb	Pending	View order	Edit configuration	Journal	Delete

Completed Jobs(1)

UID	Name	Status		
20151223022925940	mhwz	Finished - Abnormal exit from crawling	Crawl order	Crawl report

At the bottom, the identifier is: org.archive.crawler:jmxport=8849,name=1,type=CrawlService,guiport=9000,host=hfst24

第一节、采集

- 创建采集任务:

```
HERIPIX Status as of 十一月. 2, 2016 01:31:59 GMT
HOLDING JOBS
New via a profile 1 jobs pending, 1 completed
Console Jobs Profiles Logs Reports Setup Help

Select profile to base new job on:
  • govsite
  • default

Identifier: org.archive.crawler:jmxport=8849,name=1,type=CrawlService,guiport=9000,h
```

第一节、采集

- 我们在name of job处填写任务名称，可以是网址名称的首拼音缩写，seeds处填写网址。点击submit job。

127.0.0.1:9000/jobs/new.jsp?job=govsite

HERITRIX Status as of 十一月. 2, 2016 01:33:03 GMT Alerts: no al
HOLDING JOBS
New crawl job 1 jobs pending, 1 completed

Console **Jobs** Profiles Logs Reports Setup Help

Create new crawl job based on profile 'govsite'

Name of new job: wsdc


Description: Default Profile

Seeds: Fill in seed URIs below, one per line. Comment lines begin with '#'.
http://sij.hefei.gov.cn/

Modules Submodules Settings Overrides **Submit job**

第一节、采集

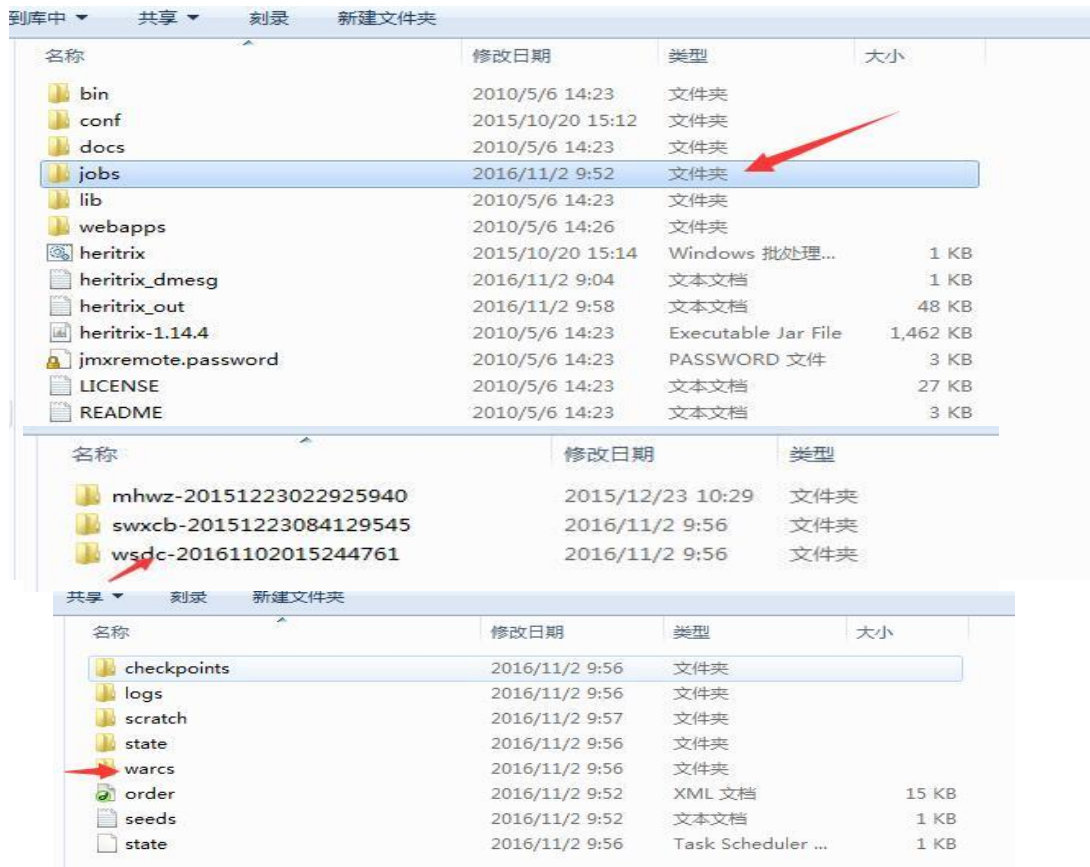
- 任务创建成功后，点击console，并点击start开始采集。

```
██████████ CRAWLING JOBS RUNNING job: wsdc
Admin Console 0 jobs pending, 2 completed 20 URIs in 3s (0/sec)
Console Jobs Profiles Logs Reports Setup Help
Crawler Status: CRAWLING JOBS | Hold
Jobs Memory
Running: wsdc 32720 KB used
0 pending, 2 completed 41112 KB current heap
Alerts: 0 (0 new) 989888 KB max heap
Job Status: RUNNING | Pause | Checkpoint | Terminate
Rates Load
0 URIs/sec (0 avg) 1 active of 50 threads
0 KB/sec (0 avg) 1 congestion ratio
Time 244 deepest queue
3s elapsed 244 average depth
48s remaining (estimated)
Totals
downloaded 20  7% 244 queued
265 total downloaded and queued
212 KB crawled (212 KB novel)
```

[Refresh](#)

第一节、采集

- 采集完成后可在job目录下找到后缀名为.warc文件。这样采集的步骤就算完成了。



第二节、发布


















- 采集后发布：先把apache解压在F盘根目录下，然后将wayback里的root.war压缩文件放置apache-tomcat的webapps目录下。

名称	修改日期	类型	大小
144共享	2016/1/23 11:58	文件夹	
apache-tomcat-6.0.44	2016/1/6 14:21	文件夹	
heritrix-gov	2016/1/21 17:37	文件夹	
temp	2016/4/16 17:10	文件夹	
tmp	2016/1/6 16:10	文件夹	
Wayback	2016/1/6 14:16	文件夹	
网事典藏文件	2016/1/23 8:49	文件夹	
新建文件夹	2015/7/19 11:51	文件夹	
apache-tomcat-6.0.44	2015/11/23 14:07	WinRAR ZIP 压缩...	5,893 KB
heritrix-gov	2015/11/23 11:27	WinRAR ZIP 压缩...	22,275 KB
jdk-8u65-windows-i586	2015/12/23 17:56	应用程序	185,579 KB
副本合肥市少年儿童图书馆-网站搜集列...	2016/1/3 10:49	Microsoft Office...	62 KB

bin	2015/5/8 20:21	文件夹	
conf	2015/5/8 20:21	文件夹	
lib	2015/5/8 20:21	文件夹	
logs	2016/11/2 10:34	文件夹	
temp	2016/11/2 10:36	文件夹	
webapps	2016/1/6 16:10	文件夹	
work	2016/1/6 16:07	文件夹	
LICENSE	2015/5/8 20:21	文件	
NOTICE	2015/5/8 20:21	文件	
RELEASE-NOTES	2015/5/8 20:21	文件	
RUNNING	2015/5/8 20:21	文本文档	

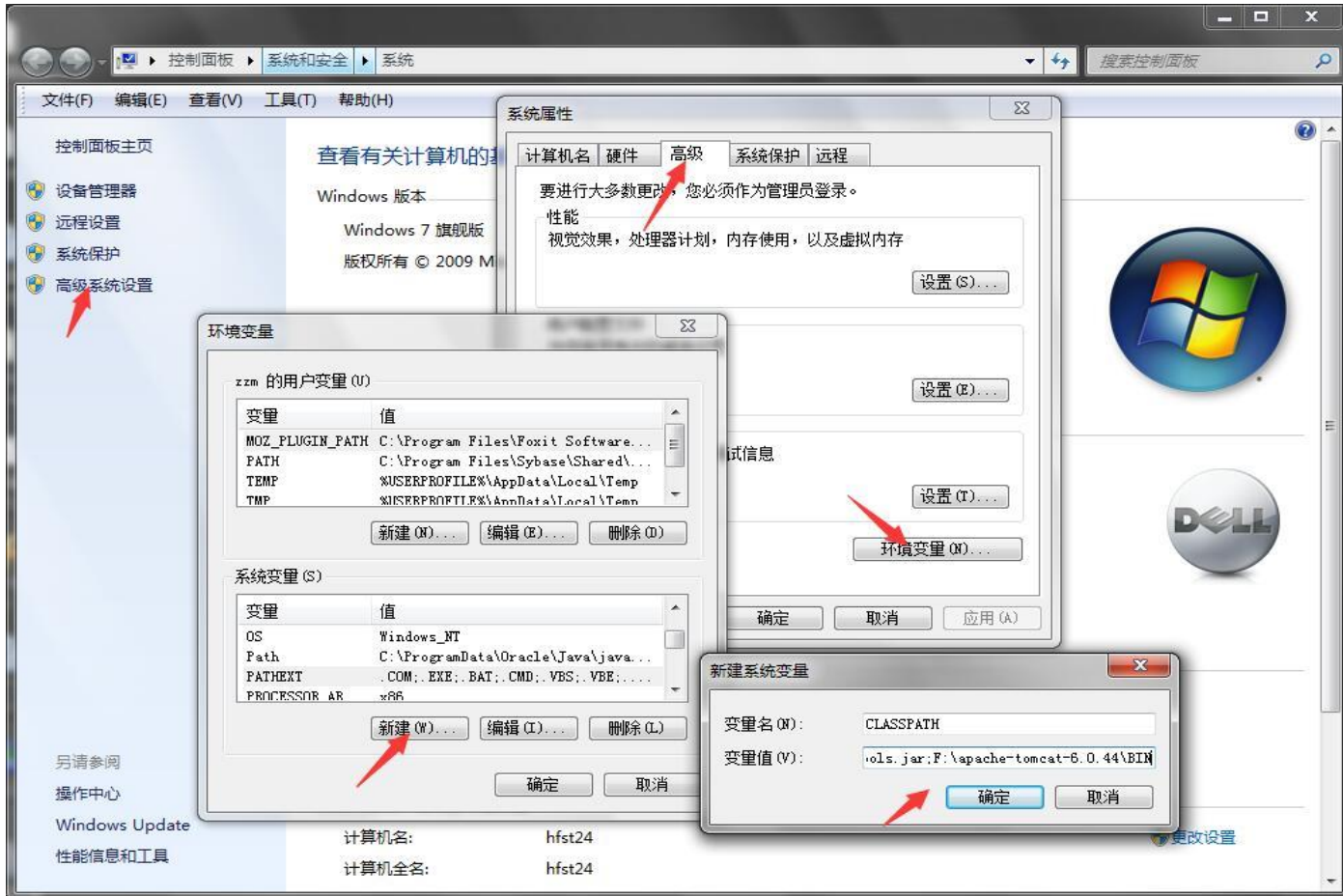
第二节、发布

- 打开bin目录下找到startup程序。

 bin	2015/5/8 20:21	文件夹	
 conf	2015/5/8 20:21	文件夹	
 lib	2015/5/8 20:21	文件夹	
 logs	2016/11/2 10:34	文件夹	
 temp	2016/11/2 10:36	文件夹	
 webapps	2016/1/6 16:10	文件夹	
 work	2016/1/6 16:07	文件夹	
 LICENSE	2015/5/8 20:21	文件	
 NOTICE	2015/5/8 20:21	文件	
 RELEASE-NOTES	2015/5/8 20:21	文件	
 RUNNING	2015/5/8 20:21	文本文档	
 setclasspath.sn	2015/5/8 20:21	SH 文件	3 KB
 shutdown	2015/5/8 20:21	Windows 批处理...	2 KB
 shutdown.sh	2015/5/8 20:21	SH 文件	2 KB
 startup	2015/5/8 20:21	Windows 批处理...	2 KB
 startup.sh	2015/5/8 20:21	SH 文件	2 KB
 tomcat-juli	2015/5/8 20:21	Executable Jar File	32 KB

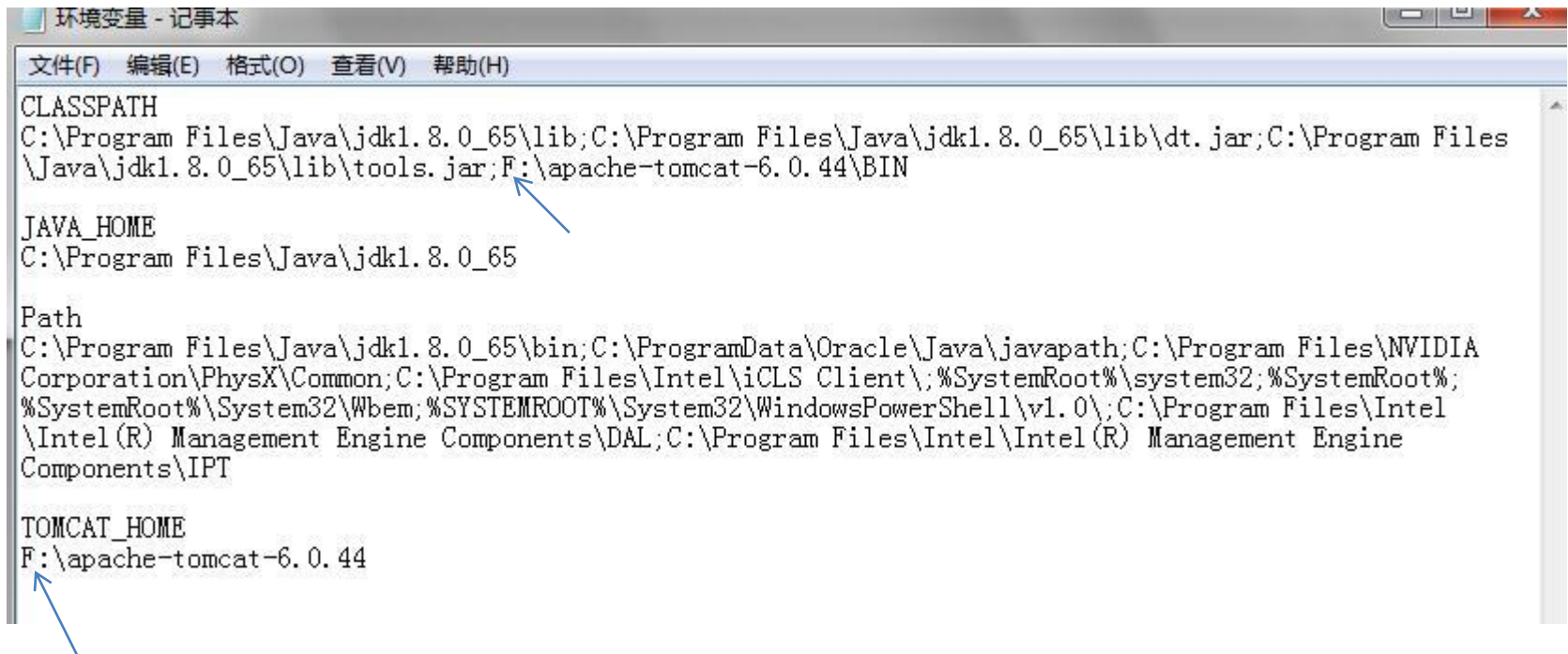
第二节、发布

- 环境变量的配置：我的电脑-属性



第二节、发布

- 环境变量（箭头所指：tomcat解压在那个盘下，就用哪个盘符）



```
环境变量 - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
CLASSPATH
C:\Program Files\Java\jdk1.8.0_65\lib;C:\Program Files\Java\jdk1.8.0_65\lib\dt.jar;C:\Program Files\Java\jdk1.8.0_65\lib\tools.jar;F:\apache-tomcat-6.0.44\BIN

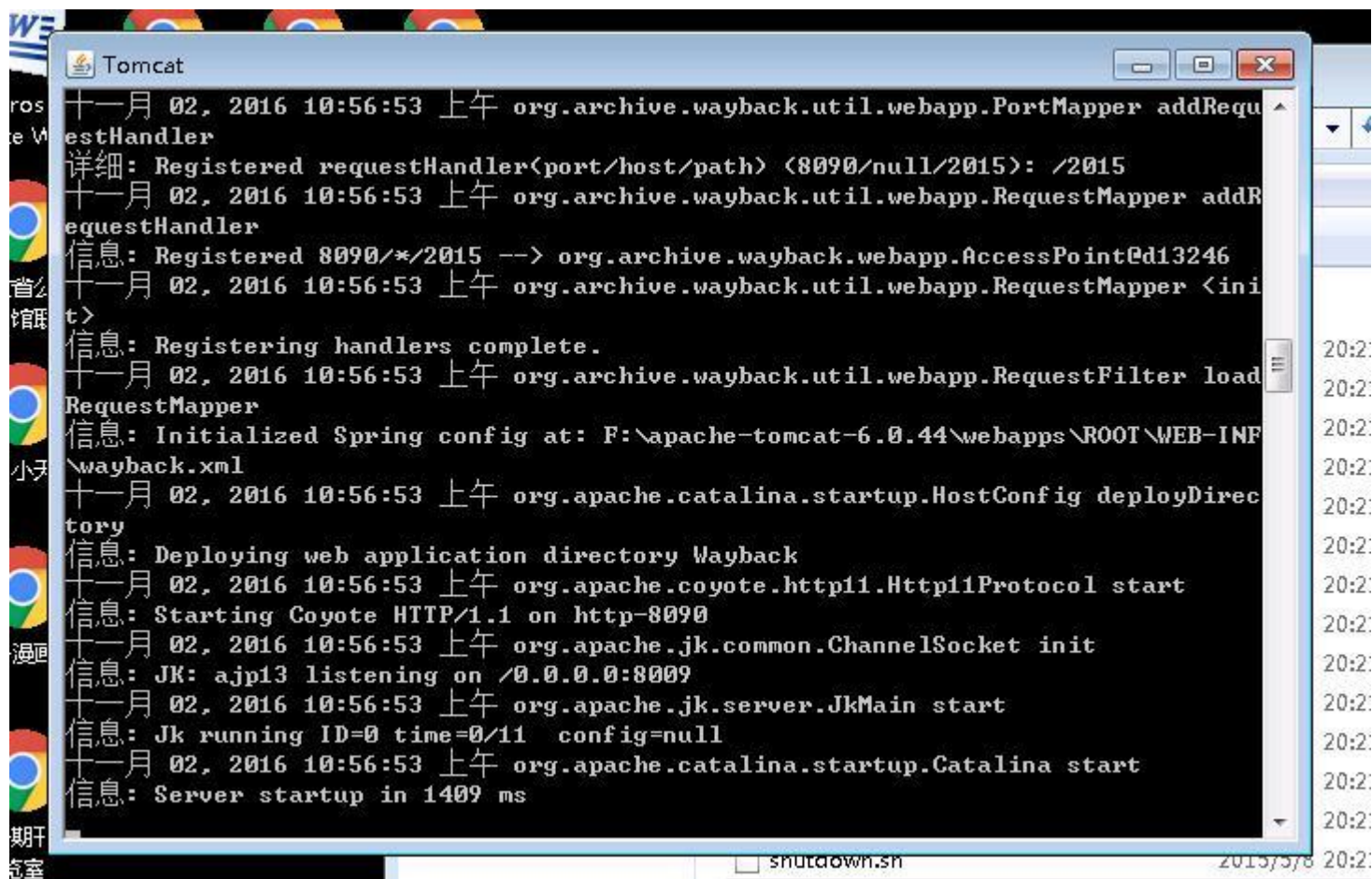
JAVA_HOME
C:\Program Files\Java\jdk1.8.0_65

Path
C:\Program Files\Java\jdk1.8.0_65\bin;C:\ProgramData\Oracle\Java\javapath;C:\Program Files\NVIDIA Corporation\PhysX\Common;C:\Program Files\Intel\iCLS Client\;%SystemRoot%\system32;%SystemRoot%;%SystemRoot%\System32\Wbem;%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\;C:\Program Files\Intel\Intel(R) Management Engine Components\DAL;C:\Program Files\Intel\Intel(R) Management Engine Components\IPT

TOMCAT_HOME
F:\apache-tomcat-6.0.44
```

第二节、发布

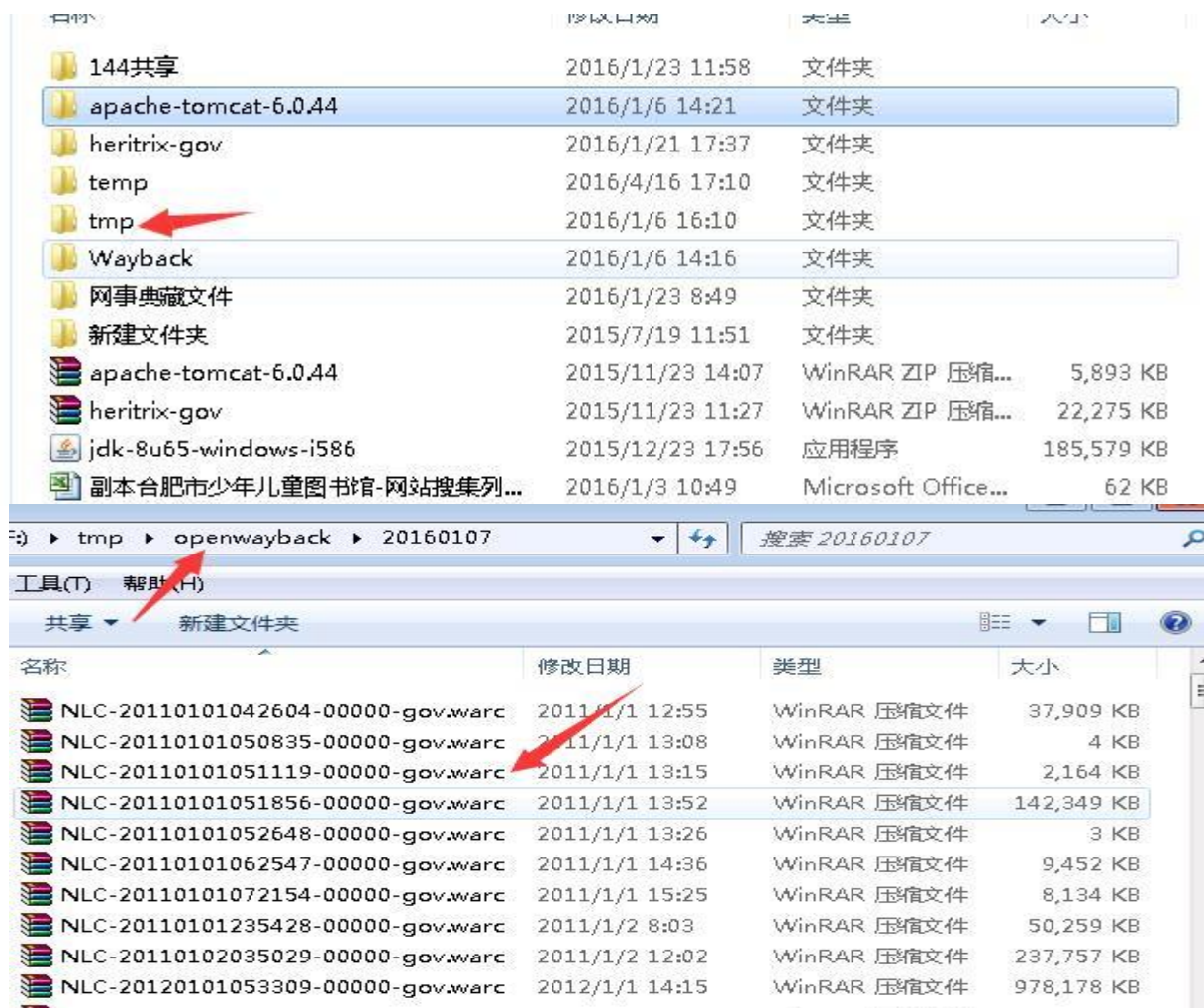
- 注意：如果一闪而过，说明环境变量没配置好。
- 点开startup后运行如下：



```
十一月 02, 2016 10:56:53 上午 org.archive.wayback.util.webapp.PortMapper addRequ
estHandler
详细: Registered requestHandler(port/host/path) (8090/null/2015): /2015
十一月 02, 2016 10:56:53 上午 org.archive.wayback.util.webapp.RequestMapper addR
equestHandler
信息: Registered 8090/*/2015 --> org.archive.wayback.webapp.AccessPoint@d13246
十一月 02, 2016 10:56:53 上午 org.archive.wayback.util.webapp.RequestMapper <ini
t>
信息: Registering handlers complete.
十一月 02, 2016 10:56:53 上午 org.archive.wayback.util.webapp.RequestFilter load
RequestMapper
信息: Initialized Spring config at: F:\apache-tomcat-6.0.44\webapps\ROOT\WEB-INF
\wayback.xml
十一月 02, 2016 10:56:53 上午 org.apache.catalina.startup.HostConfig deployDirec
tory
信息: Deploying web application directory Wayback
十一月 02, 2016 10:56:53 上午 org.apache.coyote.http11.Http11Protocol start
信息: Starting Coyote HTTP/1.1 on http-8090
十一月 02, 2016 10:56:53 上午 org.apache.jk.common.ChannelSocket init
信息: JK: ajp13 listening on /0.0.0.0:8009
十一月 02, 2016 10:56:53 上午 org.apache.jk.server.JkMain start
信息: Jk running ID=0 time=0/11 config=null
十一月 02, 2016 10:56:53 上午 org.apache.catalina.startup.Catalina start
信息: Server startup in 1409 ms
```


第二节、发布

- 将之前我们搜集到的warc文件放到盘根目录下的tmp/openwayback。



第二节、发布

- 重新启动startup程序后，我们用浏览器访问，本地电脑用<http://127.0.0.1:8090>,局域网内电脑可以通过访问ip，此时该演示的电脑IP为192.168.0.144，地址即为<http://192.168.0.144:8090>.



Enter Web Address: All

You seem to be accessing this OpenWayback via an incorrect URL. Please try one of the following AccessPoints:

[015](#) 

This is OpenWayback. Any URL in ARC or WARC files accessible to this service can be searched above.

[Home](#) | [Help](#)



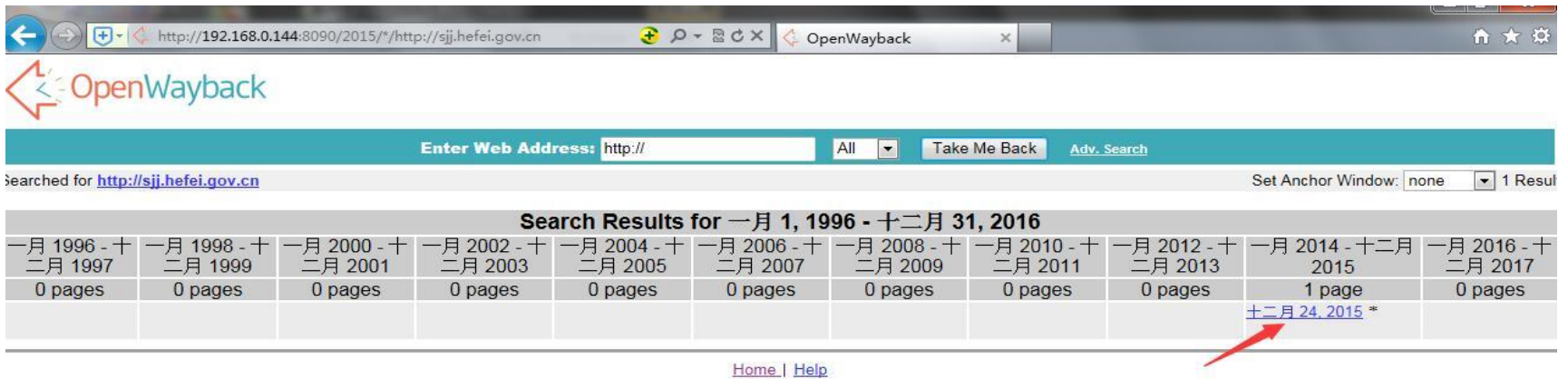
Enter Web Address: All [Adv. Search](#)

This is OpenWayback. Any URL in ARC or WARC files accessible to this service can be searched above.

[Home](#) | [Help](#)

第二节、发布

- 将之前搜集的网址输入进去，并点击。



The screenshot shows the OpenWayback interface. The search bar contains the URL `http://`. The search results are displayed in a table format, showing the number of pages available for various time periods. A red arrow points to the entry for December 24, 2015, which shows 1 page available.

Search Results for 一月 1, 1996 - 十二月 31, 2016										
一月 1996 - 十二月 1997	一月 1998 - 十二月 1999	一月 2000 - 十二月 2001	一月 2002 - 十二月 2003	一月 2004 - 十二月 2005	一月 2006 - 十二月 2007	一月 2008 - 十二月 2009	一月 2010 - 十二月 2011	一月 2012 - 十二月 2013	一月 2014 - 十二月 2015	一月 2016 - 十二月 2017
0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	1 page	0 pages
									十二月 24, 2015 *	

[Home](#) | [Help](#)

第二节、发布

- 至此，我们发布的任务也就结束了。






第三节、唯一标识符注册

- 唯一标识符注册
- 利用省馆的唯一标识符注册系统。下载模版，并按照标识符规则填写好，再批量进行注册。

您的当前位置：元数据注册 > 批量元数据注册

批量元数据注册

模板下载  excel模板  access模板  xml模版

数据来源系统

导入模板 未选择任何文件

A	B	C	D	E	F	G	H	I	J	K	L	M	N
资源种类编号	格式编号	系统号	颗粒度K1	K1值	颗粒度K2	K2值	颗粒度K3	K3值	MARC记录00	题名(资源名称作者)	ISBN	ISSN	
T8	F3		1							合肥市委宣传部			
T8	F3		2							合肥市统计局			
T8	F3		3							合肥市审计局			




语种	出版者	出版时间	关联	来源	描述信息	扩展字段1	扩展字段2	扩展字段3	扩展字段4
chi				合肥市少年	合肥市委宣传部网站			合肥市委宣传部	
chi				合肥市少年	合肥市统计局网站			合肥市统计局	
chi				合肥市少年	合肥市审计局网站			合肥市审计局	

第三节、唯一标识符注册

- 把所有采集的网站按照模版填写好，就可以导入生成cdoi

您的当前位置：元数据注册>批量元数据注册

批量元数据注册

模板下载  excel模板  access模板  xml模版

数据来源系统

导入模板 未选择任何文件

第三节、唯一标识符注册

- 在元数据维护中可以看到批量注册的结果

系统菜单

- 元数据注册
- 元数据维护
- >元数据维护
- >元数据维护(批量)
- >URL维护(批量)
- >元数据导出
- >批量处理结果
- >元数据批量删除

您的当前位置：元数据维护>批量处理结果

信息检索

处理类型 上传时间

查询

批量处理结果信息一览

总计7条记录

序号	处理批次号	上传帐号	处理类型	处理状态	处理结果描述	上传时间	处理时间	基本操作
1	2016012210080021	hfse	URL批量维护	已处理	维护成功数量为(201)维护失败数量为(0)	2016-01-22 11:38:44	2016-01-22 11:39:50	
2	2016012210020044	hfse	元数据批量注册	已处理	注册成功数量为(3)注册失败数量为(0)	2016-01-22 11:33:12	2016-01-22 11:33:47	
3	2016012114260042	hfse	元数据批量注册	已处理	注册成功数量为(200)注册失败数量为(0)	2016-01-21 15:54:49	2016-01-21 15:57:31	
4	2016011410470026	hfse	元数据批量注册	已处理	注册成功数量为(3)注册失败数量为(0)	2016-01-14 12:14:44	2016-01-14 12:15:32	
5	2016011315560001	hfse	URL批量维护	已处理	维护成功数量为(2)维护失败数量为(0)	2016-01-13 17:22:16	2016-01-13 17:24:12	
6	2016011315350024	hfse	元数据批量注册	已处理	注册成功数量为(2)注册失败数量为(0)	2016-01-13 17:02:31	2016-01-13 17:03:12	
7	2016011315260022	hfse	元数据批量注册	已处理	注册成功数量为(1)注册失败数量为(2)	2016-01-13 16:53:27	2016-01-13 16:54:12	

第三节、唯一标识符注册

安徽省图书馆唯一标识符系统

用户手册 本机构信息 帮助说明 修改密码 退出系统

hfse [机构管理员 / 合肥市少儿图书馆] 上午好!

系统菜单

- 元数据注册
- 元数据维护
- >元数据维护
- >元数据维护(批量)
- >URL维护(批量)
- >元数据导出
- >批量处理结果
- >元数据批量删除

您的当前位置: 元数据维护>元数据导出

信息检索

唯一标识符:

机构名称: 合肥市少儿图书馆 数据来源系统: WSDC

创建时间(始): 创建时间(止):

交易流水号:

检索字段: 选择查询条件 精确

元数据导出

批量导出 全部导出

总计203条记录 下一页 末页 1 / 7页 GO

<input type="checkbox"/>	唯一标识符	题名(资源名称)	文件格式	资源类型	基本操作
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.203	安徽肥西桃花工业园区	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.202	肥东县党的群众路线教育实践活动网	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.201	肥西县市场监督管理局	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.200	巢湖共青团	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.199	合肥共青团	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.198	共青团肥西县委员会	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.197	肥西县文化广电新闻出版局	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.196	肥东县畜牧水产局	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.195	肥西县官亭镇政府	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.194	肥西县重点工程建设管理局	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.193	肥西县丰乐镇人民政府	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.192	肥西县花岗镇人民政府	ASP	网页采集	
<input type="checkbox"/>	108.ndlc.18.3401039031010002/T8F3.191	肥西县紫蓬镇人民政府	ASP	网页采集	

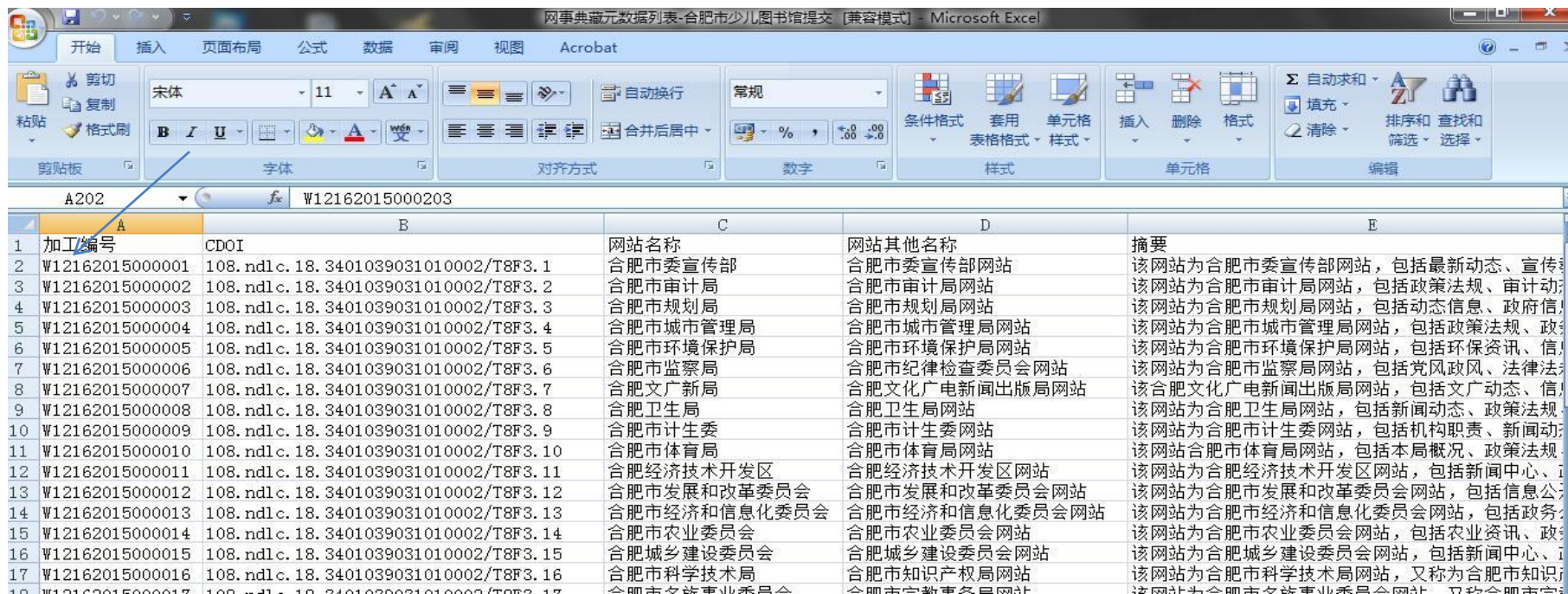
第三节、唯一标识符注册

- 做完唯一标识符注册后需要进行url维护。
- 和注册类似，完成后唯一标识符的注册也就完成了。



第四节、元数据及质检报告

- 制作元数据
- 按照之前数据采集发布及唯一标识符注册的内容填写EXCEL表格。



The screenshot shows a Microsoft Excel spreadsheet with the following data:

加工/编号	CD01	网站名称	网站其他名称	摘要
W12162015000001	108.ndlc.18.3401039031010002/T8F3.1	合肥市委宣传部	合肥市委宣传部网站	该网站为合肥市委宣传部网站,包括最新动态、宣传动
W12162015000002	108.ndlc.18.3401039031010002/T8F3.2	合肥市审计局	合肥市审计局网站	该网站为合肥市审计局网站,包括政策法规、审计动
W12162015000003	108.ndlc.18.3401039031010002/T8F3.3	合肥市规划局	合肥市规划局网站	该网站为合肥市规划局网站,包括动态信息、政府信
W12162015000004	108.ndlc.18.3401039031010002/T8F3.4	合肥市城市管理局	合肥市城市管理局网站	该网站为合肥市城市管理局网站,包括政策法规、政
W12162015000005	108.ndlc.18.3401039031010002/T8F3.5	合肥市环境保护局	合肥市环境保护局网站	该网站为合肥市环境保护局网站,包括环保资讯、信
W12162015000006	108.ndlc.18.3401039031010002/T8F3.6	合肥市监察局	合肥市纪律检查委员会网站	该网站为合肥市监察局网站,包括党风政风、法律法
W12162015000007	108.ndlc.18.3401039031010002/T8F3.7	合肥文广新局	合肥文化广电新闻出版局网站	该合肥文化广电新闻出版局网站,包括文广动态、信
W12162015000008	108.ndlc.18.3401039031010002/T8F3.8	合肥卫生局	合肥卫生局网站	该网站为合肥卫生局网站,包括新闻动态、政策法规
W12162015000009	108.ndlc.18.3401039031010002/T8F3.9	合肥市计生委	合肥市计生委网站	该网站为合肥市计生委网站,包括机构职责、新闻动
W12162015000010	108.ndlc.18.3401039031010002/T8F3.10	合肥市体育局	合肥市体育局网站	该网站合肥市体育局网站,包括本局概况、政策法规
W12162015000011	108.ndlc.18.3401039031010002/T8F3.11	合肥经济技术开发区	合肥经济技术开发区网站	该网站为合肥经济技术开发区网站,包括新闻中心、
W12162015000012	108.ndlc.18.3401039031010002/T8F3.12	合肥市发展和改革委员会	合肥市发展和改革委员会网站	该网站为合肥市发展和改革委员会网站,包括信息公
W12162015000013	108.ndlc.18.3401039031010002/T8F3.13	合肥市经济和信息化委员会	合肥市经济和信息化委员会网站	该网站为合肥市经济和信息化委员会网站,包括政务
W12162015000014	108.ndlc.18.3401039031010002/T8F3.14	合肥市农业委员会	合肥市农业委员会网站	该网站为合肥市农业委员会网站,包括农业资讯、政
W12162015000015	108.ndlc.18.3401039031010002/T8F3.15	合肥城乡建设委员会	合肥城乡建设委员会网站	该网站为合肥城乡建设委员会网站,包括新闻中心、
W12162015000016	108.ndlc.18.3401039031010002/T8F3.16	合肥市科学技术局	合肥市知识产权局网站	该网站为合肥市科学技术局网站,又称为合肥市知识
W12162015000017	108.ndlc.18.3401039031010002/T8F3.17	合肥市各民主党派委员会	合肥市各民主党派网站	该网站为合肥市各民主党派委员会网站,又称合肥市

第四节、元数据及质检报告

关键词	资源类型	内容形式	媒体类型	语种	保存格式	机构名称	行政级别	关联	访问方式	采集日期
合肥; 宣传; 市委	网站	多种内容形式	电子	chi	WARC	合肥市委宣传部	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 审计	网站	多种内容形式	电子	chi	WARC	合肥市审计局	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 规划局	网站	多种内容形式	电子	chi	WARC	合肥市规划局	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 城管; 城市管理	网站	多种内容形式	电子	chi	WARC	合肥市城市管理局	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 环保	网站	多种内容形式	电子	chi	WARC	合肥市环境保护局	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 监察; 纪律	网站	多种内容形式	电子	chi	WARC	合肥市监察局	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 文广; 广电; 文化; 广电	网站	多种内容形式	电子	chi	WARC	合肥文广新局	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 卫生	网站	多种内容形式	电子	chi	WARC	合肥卫生局	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 计生委	网站	多种内容形式	电子	chi	WARC	合肥市计生委	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 体育	网站	多种内容形式	电子	chi	WARC	合肥市体育局	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 经济; 开发区	网站	多种内容形式	电子	chi	WARC	合肥经济技术开发区	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 发展; 改革; 委员会	网站	多种内容形式	电子	chi	WARC	合肥市发展和改革委员会	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-
合肥; 经济; 信息; 委员会	网站	多种内容形式	电子	chi	WARC	合肥市经济和信息化委员会	县(区)级及以下		合肥市少年儿童图书馆局域网访问	2015-12-

P	Q	R	S	T	U	V	W
采集日期	发布日期	采集地址	发布地址	附注	数据提交单位	所属任务年份	
2015-12-30	2016-01-09	http://swxcb.hefei.gov.cn/	http://192.168.0.144:8090/2015/20151230		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://s.ji.hefei.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://www.hfsghj.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://www.hfsr.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://www.hfepb.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://www.hfsjw.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://swhj.hefei.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://swsj.hefei.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://sjsj.hefei.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://stvj.hefei.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://www.hetda.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://www.hfdpc.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://www.hfgj.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2015-12-24	2016-01-09	http://www.hfac.gov.cn/	http://192.168.0.144:8090/2015/20151224		合肥市少年儿童图书馆	2015	
2016-01-13	2016-01-14	http://www.hfjs.gov.cn/jw	http://192.168.0.144:8090/2015/20160113		合肥市少年儿童图书馆	2015	

第四节、元数据及质检报告

- 第三方质检报告。到此我们网事典藏的项目就算是完成了。

推广工程数字资源联合建设

项目质检报告

2016年1月27日

项目名称	网事典藏项目数据建设		
建设单位	合肥市少年儿童图书馆	建设数量	200条
质检单位	合肥东联信息科技有限公司	抽检数量	100条
提交时间	2016-1-22	完成时间	2016-1-27
质检结果	合格		
<p>一、总体说明</p> <p>本次质检为抽检，由合肥东联信息科技有限公司员工，对其中100条网事典藏的元数据进行检查，主要检查了加工编号、CD01、网站名称、机构名称、采集地址和发布地址等这些字段。</p> <p>二、质检过程</p> <p>本项目质检主要通过目测+比对的方法对元数据进行抽查，主要对采集地址和发布地址进行点击访问，查看发布地址的网页是否能够正常访问，并根据网站信息对其他字段进行逐一检查。</p> <p>三、主要问题</p> <p>本次质检共抽检了元数据100条，主要检查其中的加工编号、CD01、机构名称、采集和发布地址，主要问题有发布的网页中有部分图片、子页面打不开。但因这些图片和子页面不属于该网站域名下，为外部链接，故不认定为错，所以抽检的合格率为100%，质检结果为合格。</p> <p>四、检查明细</p> <p>抽检了其中100条元数据，抽检结果为合格。</p>			

第二章：项目重难点和完成技巧

- 1、做好前期规划，设置时间节点，合理进行任务分解，按时按量完成阶段性任务，预留反应时间。
- 2、遇到问题及时请教省图或国图，保证项目始终处在正确进展中，避免返工。
- 3、一定要网站列表通过省图或国图审核通过后，再开始采集，避免无用功。
- 4、项目花时间最多的环节是网站采集，所以我们采用了多台电脑同时进行采集。
- 5、采集发布的环境配置可能会出现如下问题：应该如何避免和解决：
 - （1）系统环境：建议在WIN7搭建。
 - （2）采集步骤：

如果网站采集过程中遇到采集缓慢，或者网站采集不到，建议更换网站进行采集。

第二章：项目重难点和完成技巧

（3）发布环节：

环境变量的配置，还有两个XML文件的参数修改问题。

（4）唯一标识符：

如果本馆没有自建，利用省馆搭建的唯一标识符系统，由省馆提供的账号和密码登录进入系统完成注册。

（5）制作元数据：

制作元数据时碰到的问题，加工编号里面的机构代码，当时国图并没有给少儿馆分，在文件中也找不到，所以国图老师临时分了1216作为我馆的机构代码。

- 6、做好数据备份，至少在本地要有两个副本，移动存储中再拷贝一份。

结束语

- 谢谢大家！
- 推荐两个群！
- 联建网事典藏项目群：365776635。
- 唯一标识符：378362263。