2015年资源联建项目培训

——网事典藏项目建设标准及方法

国家图书馆 数字资源部

2015年12月

主要内容

一
- 网络资源采集与存档

二
- 资源联建网事典藏项目

网络资源采集与存档

1 网络存档简介

- 网络资源采集与存档（Web Archiving）

  网站的一种镜像

  利用爬虫进行

  有特定的时间

  看起来像是真的网站一样

- 利用爬虫在特定的时间对网站（网络资源）进行抓取，制作镜像；用户访问存档资源时像是真的网站一样。

网络资源采集与存档

2 国内外相关项目

- 为了保存互联网上的信息，从20 世纪90年代中期开始，就有很多国家开展了网络资源存档项目，采集并保存了大量的互联网资源。

国家	机构	项目	开始时间
美国	互联网档案馆 Internet Archive	时光机 Wayback	1996
		存档它 Archive-It	
	美国国会图书馆	雅典娜 Minerva	2000
法国	法国国家图书馆	BnF Internet Archives	2002
英国	英国国家图书馆	UK Web Archive	2004
	英国国家档案馆	UK Government Web Archive	1997

网络资源采集与存档

2 国内外相关项目

国家	机构	项目	开始时间
澳大利亚	澳大利亚国家图书馆	潘多拉 PANDORA	1996
芬兰	芬兰国家图书馆	Finnish Web Archive	2006
瑞典	瑞典国家图书馆	Kulturarw[3]	1997
冰岛	冰岛国家与大学图书馆	Icelandic Web Archive	1996
加拿大	加拿大国家图书馆与档案馆	Government of Canada Web Archive	2005
挪威	挪威国家档案馆	Web Archive Norway	2001

网络资源采集与存档

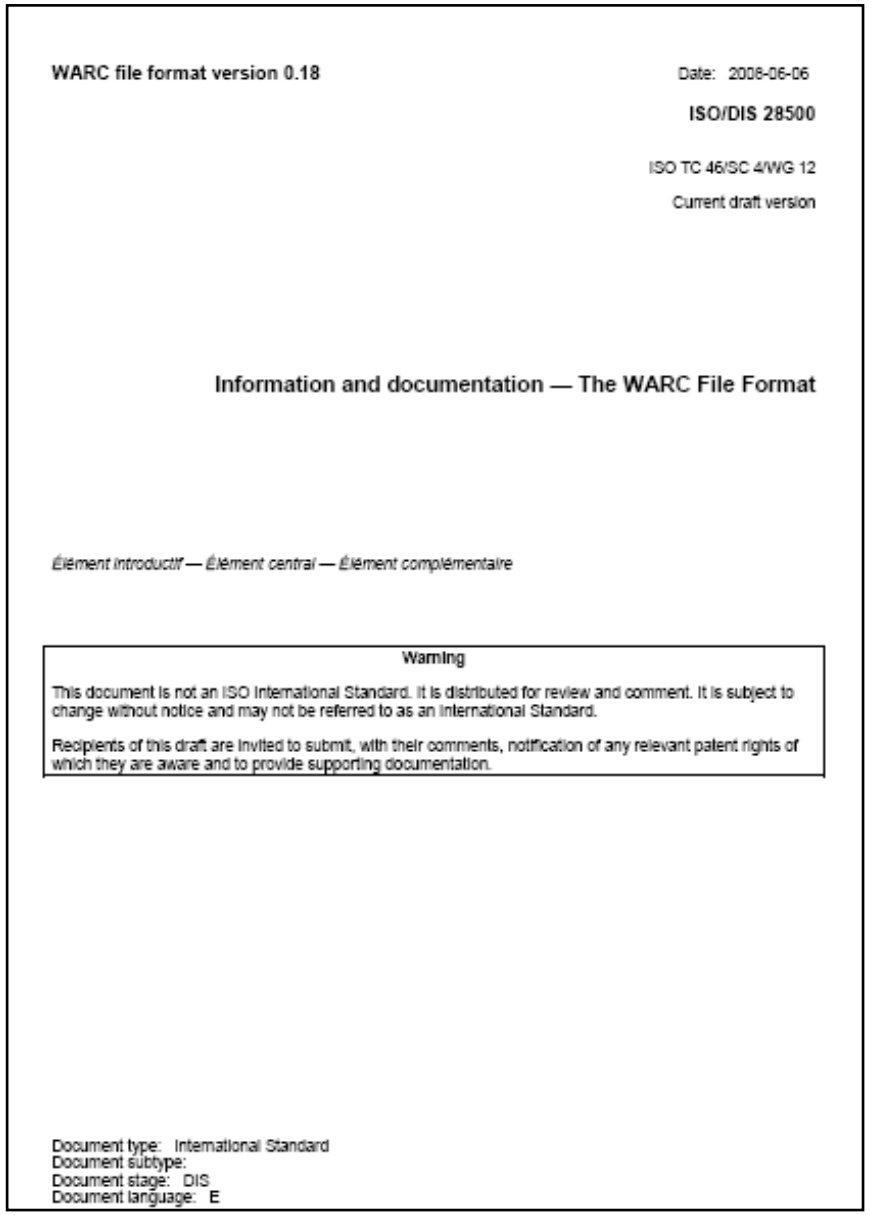
2 国内外相关项目

国家	机构	项目	开始时间
新西兰	新西兰国家图书馆	New Zealand Web Archive	1999
日本	日本国立国会图书馆	WARP	2004
韩国	韩国国家图书馆	绿洲　OASIS	2000
中国	中国国家图书馆	WICP	2003
	北京大学	中国Web信息博物馆　Web Informall	2002-2012
	台湾图书馆	Web Archive Taiwan	1996
	台湾大学图书馆	台大图书馆网站典藏库	2006

网络资源采集与存档

3 网络存档的标准

- WARC 格式
- WARC = Web ARChive file format
- 网络信息资源的保存格式
- 大文件格式，内嵌元数据的对象格式
- ISO 28500：2009



WARC file format version 0.18

Date: 2008-06-06

ISO/DIS 28500

ISO TC 46/SC 4/WG 12

Current draft version

Information and documentation — The WARC File Format

Élément introductif — Élément central — Élément complémentaire

Warning

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Document type: International Standard
Document subtype:
Document stage: DIS
Document language: E

网络资源采集与存档

4 国家图书馆的网络存档工作

资源联建网事典藏项目

一	• 建设内容
二	• 工作流程
三	• 成果提交
四	• 元数据著录规则
五	• 网页采集和网页发布软件

一、建设内容

网事典藏项目建设初期先以政府网站的采集和存档为重点，主要采集反映所在行政区域的政治、经济、文化发展等信息的政府网站，并将采集到的网站进行编目和发布。各馆原则上只采集行政上从属于本地区的政府网站。

一、建设内容

注意 与政府公开信息项目的区别

采集和著录的对象不同

政府公开信息项目	网事典藏项目
网站中每一条具体的政府信息	整个政府网站

项目采集的目的不同

政府公开信息项目	网事典藏项目
是为了获取网页中想要的内容为主，对内容进行整合，以应用服务为主。	保存网页原貌，以存档为主。同一网站多次采集。

二、工作流程

1. 采集准备
2. 资源采集
3. 数据发布
4. 元数据制作
5. 数据验收
6. 数据维护和长期保存

二、工作流程

1 采集准备

- 将需要采集的政府网站网址（URL地址）整理成采集列表（**excel表格**），表头如下：

政府网站采集列表

序号	网站名称	网站域名	备注

- 市馆提交给省馆初审，省馆初审后，连同初审意见一同提给交国家图书馆审核，由国家图书馆出具审核意见。

二、工作流程

2 资源采集

根据采集列表，利用网络采集软件，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域名内的全部内容，但不包括论坛等需链接后台数据库的内容。所采集的文档格式遵循WARC1.0标准，不含病毒、垃圾文件及采集列表外的其他信息。每个网站单独采集。

采集结果：WARC文档

二、工作流程

3 数据发布

- 将采集到的文档（WARC文档）数据**进行索引**后发布。
- 保证页面内容都能正常打开，且与原网站保持一致。
- 数据需要在**推广工程专用网络**内发布。

二、工作流程

3 数据发布


http://192.168.182.61:8089/2014zy/*/http://www.mfa.gov.cn/mfa_chn/
互联网时光机
文件(F) 编辑(E) 查看(V) 收藏夹(A) 工具(T) 帮助(H)
输入需检索的URL(输入检索地址:) http://www.mfa.gov.cn/mfa_c 全部 检索 高级检索(高级检索)
检索URL： http://www.mfa.gov.cn/mfa_chn/ Set Anchor Window: none 1 Result
检索结果： 一月 1, 1996 - 十二月 31, 2014
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
0 pages | 0 pages | 0 pages | 0 pages | 0 pages | 0 pages | 0 pages | 0 pages | 0 pages | 1 page
一月 9, 2014 *
京ICP备05014420号 电话:(+86 10)88545587-805 中国国家图书馆版权所有，中国事典网站
中国事典中的存档资源目前只提供国家图书馆馆内访问，暂不提供互联网服务。
中心主页 | 中国事典主页
100%

二、工作流程

3 数据发布


http://192.168.182.61:8089/2014zy/20140109082352/http://www.mfa.gov.cn/mfa_chn/
中华人民共和国外交部
文件(F) 编辑(E) 查看(V) 收藏夹(A) 工具(T) 帮助(H)
中国事典
http://www.mfa.gov.cn/mfa_chn/
查找
1次采集
十二月 一月 二月
9
2013 2014 2015
关闭
帮助
中华人民共和国外交部
中文繁体 | English | Français | Русский | Español
请输入关键字
搜索
高级
此地址为元数据中的“发布地址”
http://10.100.1.123:8089/2014zy/20140109082352/http://www.mfa.gov.cn/mfa_chn/
AFRICAN UNION
1 2 3 4
王毅：让非盟会议中心这座中非友好丰碑
始终屹立在中非人民心中
重要新闻
更多>>
王毅：中国在中东地区发挥的政治作用只会越来越多
王毅：日本领导人应尊重人类良知和国际公理的底线
王毅：希望中美在亚太形成良性互动
王毅：泛非主义是非洲的方向和时代的潮流
王毅与加纳外长特塔赫举行会谈
王毅：稳定和发展符合南苏丹各民族的根本利益
王毅：应落实好六国与伊朗核协议
王毅：希望埃及重新作为地区大国发挥作用
习近平主席特使姜伟新将出席坦桑尼亚桑给巴尔革命胜利50周年庆典
保加利亚总统普列夫内利埃夫将访华
王毅：中国是非洲和平安全事务的积极参与者
吉布提总统盖莱会见王毅
王毅与吉布提外长优素福举行会谈
王毅：让非盟会议中心这座中非友好丰碑始终屹立在中非人民心中
埃塞俄比亚总统穆拉图会见王毅
埃塞俄比亚总理海尔马里亚姆会见王毅
100%

二、工作流程

4 元数据制作

- 《推广工程数字资源联合建设政府网站元数据著录规则》
- 每个采集结果对应一条完整的元数据。
- 需要在唯一标识符系统中注册**CDOI。**
- 将元数据制作成**excel表。**（作为成果提交）

二、工作流程

5 数据验收

各馆在规定日期前，向国家图书馆提交已由第三方机构初检合格的全部数字资源。经国家图书馆终验合格后，提交成品数据。

- 元数据审校：对编目完整的元数据按照《著录规则》进行审校，保证各字段的准确、完整。
- 对象数据审校：通过点击的方式进行查验，保证页面内容都能正常打开，且与原网站保持一致。
- 数据本馆审校合格后交**第三方进行验收**，验收不合格需要修改或重采，直到验收合格。（验收报告作为成果提交）

二、工作流程

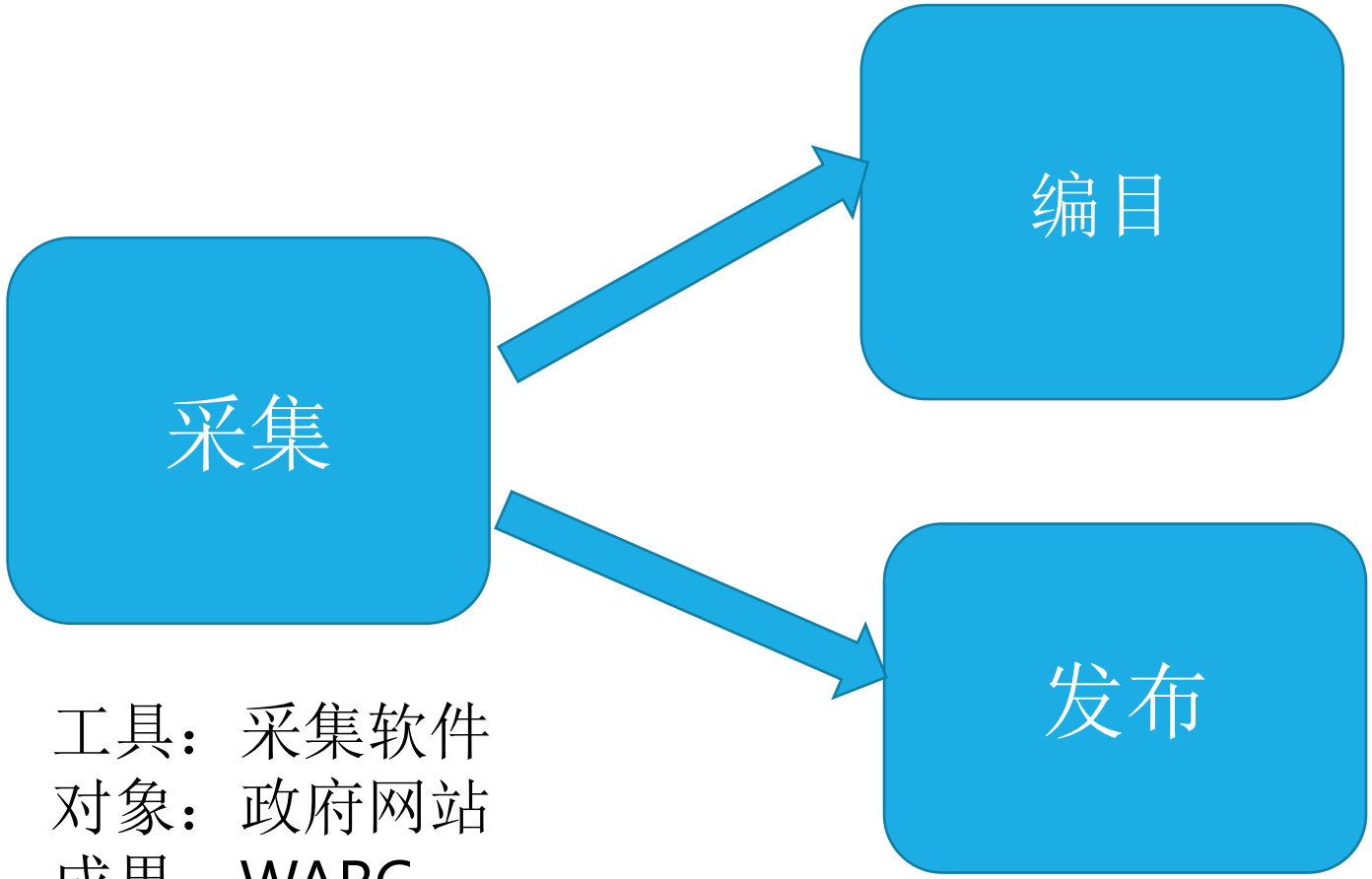
6 数据维护和长期保存

- 各馆负责对本机构制作及发布的信息及其发布网站进行长期维护，保障数据准确无误，显示正常，同时做好数据备份与长期保存工作。

三、成果提交

- 元数据：政府网站的元数据以excel表格方式提交。
- 对象数据：采集的政府网站需要在推广工程专用网络内发布，为用户提供服务。
- 第三方质检报告。

小节



工具：采集软件
对象：政府网站
成果：WARC

政府网站
元数据著录规则

工具：数据发布软件
对象：已采集的数据
成果：网站

四、元数据著录规则

《推广工程数字资源联合建设政府网站元数据著录规则》

- 著录对象。著录对象为存档的政府网站。以单次存档的政府网站为一个著录单位。如果一个政府网站具有多个主页域名，著录时作为一个对象著录。
- 著录要求。对采集的政府网站进行编目加工，要求参照著录规则进行编目。元数据以EXCEL文件形式提交。

四、元数据著录规则

术语	必备性	著录内容
加工编号	必备	著录元数据的一个明确标识，定长为15位。具体组成是：资源类型代码（1位，网站：W）、采集机构代码（4位，数值取自机构代码，2-5字符位）、采集年（4位，6-9字符位）、流水号（6位，10-15字符位）。流水号应顺序排列，不同存档资源流水号不可重复。 例如：W0100 2015 000001
CDOI	必备	著录所采集网站的唯一标识号。
网站名称	必备	著录网站名称。信息源取自网站页面首页源代码中的<title>。若<title>为空，或不反映网站内容，可用网站其他位置明显反映网站内容的名称。
网站其他名称	必备	统一著录“××网站”，对网站名称进行解释说明。例如：朝阳区人民政府网站著录的网站其他名称为“北京市朝阳区人民政府网站”。
摘要	必备	著录网站内容的总结概括性文字。摘要字数要求200字以内。
关键词	必备	著录体现网站主要内容的名词或名词短语。如有多个关键词，以半角分号间隔。
资源类型	必备	著录所保存资源的类型。统一著录为“网站”。

四、元数据著录规则

术语	必备性	著录内容
内容形式	必备	著录内容形式及内容限定。参考国家标准GB/T 3469—2013《信息资源的内容形式和媒体类型标识》取值。
媒体类型	必备	著录用以承载资源内容的载体类别。参考国家标准GB/T 3469—2013《信息资源的内容形式和媒体类型标识》取值。网络信息保存资源媒体类型统一著录为“电子”。
语种	必备	著录网站的3位语种代码，可参考《新版中国机读目录格式使用手册》。如有多个语种，以半角分号间隔。
保存格式	必备	著录所采集的网站资源存档格式。统一著录为“WARC”。
机构名称	必备	著录网站的所属机构名称。著录时应以通用性、惯用性为选取原则如网站中出现多个不同的名称，选择网站最显著位置的名称。
行政级别	有则必备	著录机构所属行政级别，取值：“中央”、“省（副省）级”、“市（地）级”、“县（区）级及以下”。

四、元数据著录规则

术语	必备性	著录内容
关联	有则必备	著录与当前资源存在某种关系的其他资源。
访问方式	必备	著录资源可以提供服务的范围，取值：互联网访问、推广工程专用网络访问等。
采集日期	必备	著录网站采集的日期。如果在审核过程中需重新采集，应对本项内容进行修改。
发布日期	必备	著录存档资源发布的日期。
采集地址	必备	著录政府网站的原始访问地址。
发布地址	必备	著录存档资源的发布地址。
附注	有则必备	凡未在其他著录项中著录而又有必要进一步补充说明的内容，均可著录于本项。
数据提交单位	必备	著录承建馆的名称。
所属任务年份	必备	著录联建工作的任务年度，2015年度数据则著录2015。

五、网页采集和网页发布软件

五、网页采集和网页发布软件

网页采集软件

- 面向归档的网络爬虫
- JAVA
- 开源
- 命名Heritrix（heiress古语词）
- Internet Archive开发
- 2003年上半年开始开发
- 2004年8月，版本1.0.0发布
- Heritrix 3.2.0(2014年1月)

https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

五、网页采集和网页发布软件

软件准备

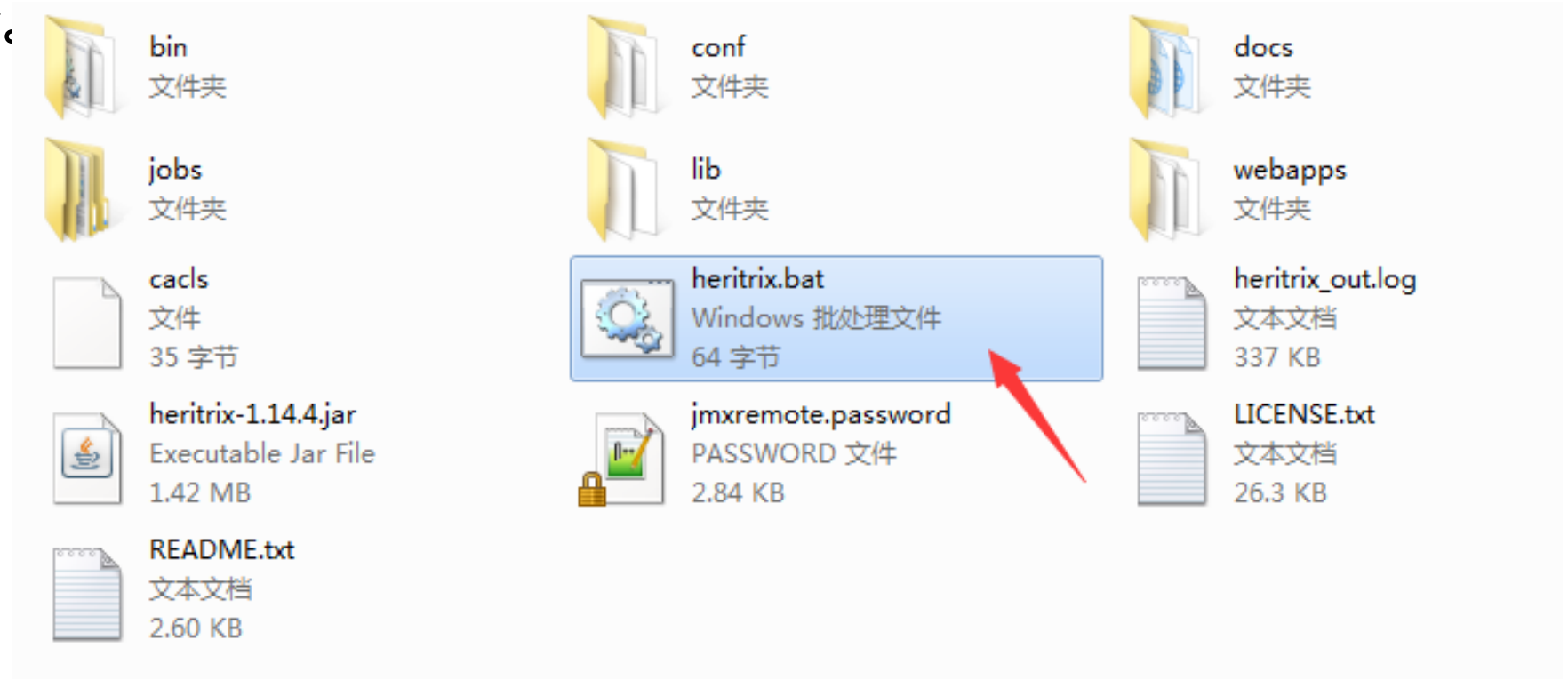
- 安装Java运行环境
- 根据系统情况安装相应版本的JDK（Java SE Development Kit ）

http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

五、网页采集和网页发布软件

1 启动Heritrix

- 在Heritrix的目录下找到“heritrix.bat”文件，双击该文件，打开heritrix服务。



五、网页采集和网页发布软件

1 启动Heritrix

- 正常启动显示的信息

```
D:\>heritrixgov22\bin\heritrix --admin=admin:admin -b 0.0.0.0 -p 9000
WARNING: It's currently not possible to run Heritrix in background
         on Windows. It was just started minimized in a new Window
         and will be shut down as soon as you log off.

2015/10/13 周二 11:11:51.10 Starting heritrix
Heritrix 1.14.4 is running.
Web console is at: http://0.0.0.0:9000
Web console login and password: admin/admin
```

- 弹出新的控制台窗口

```
03:11:52.070 EVENT  Starting Jetty/4.2.23
03:11:52.166 WARN!! Delete existing temp dir C:\Users\ADMINI~1\AppData\Local\Tem
p\2\Jetty_0_0_0_0_9000__ for WebApplicationContext[/,jar:file:/D:/heritrixgov22/
webapps/admin.war!/]
03:11:52.302 EVENT  Started WebApplicationContext[/,Heritrix Console]
03:11:52.425 EVENT  Started SocketListener on 0.0.0.0:9000
03:11:52.426 EVENT  Started org.mortbay.jetty.Server@4e515669
2015-10-13 03:11:52.962 信息 thread-1 org.archive.crawler.Heritrix.postRegister(
) org.archive.crawler:guiport=9000,host=WIN-LS216LHR4J3,jmxport=8849,name=Heritr
ix,type=CrawlService registered to MBeanServerId=WIN-LS216LHR4J3_1444705911399,
SpecificationVersion=1.4, ImplementationVersion=1.8.0_60-b27, SpecificationVendo
r=Oracle Corporation
Heritrix version: 1.14.4
```

五、网页采集和网页发布软件

启动Heritrix（问题）

- 启动时可能会有如下提示

```
D:\>heritrixgov22\bin\heritrix --admin=admin:admin -b 0.0.0.0 -p 9002
WARNING: It's currently not possible to run Heritrix in background
         on Windows. It was just started minimized in a new Window
         and will be shut down as soon as you log off.

2015/10/13 周二 11:05:08.79 Starting heritrix

Heritrix failed to start properly. Possible causes:

- Login and password have not been specified (see --admin switch)
- another program uses the port for the web UI (8080 by default)
  (e.g. another Heritrix instance)
- JMX password file is missing or permissions not set correctly

Do you want to try to fix the permissions (Y/N)?
```

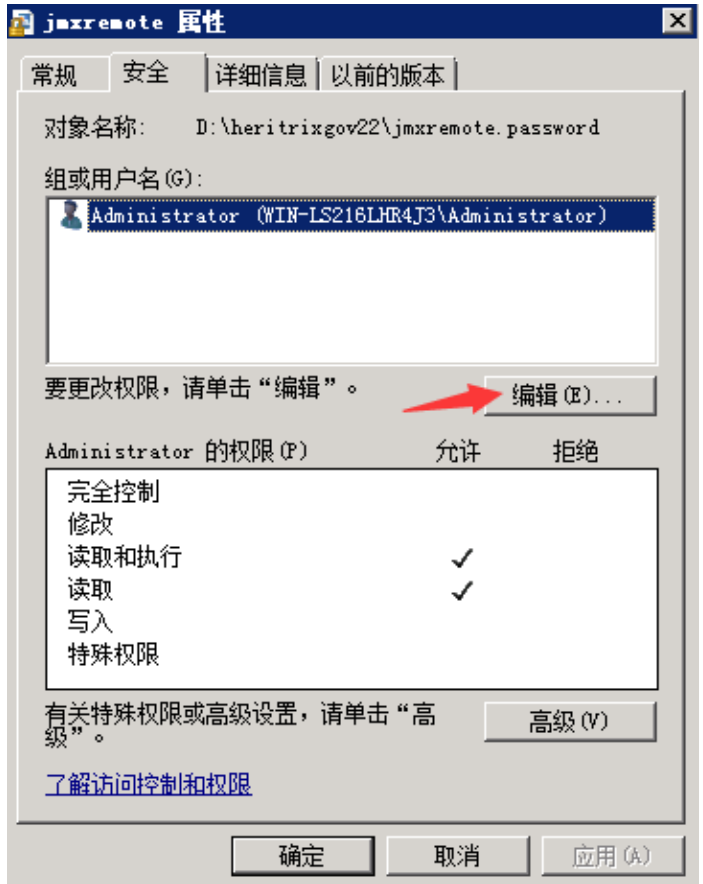
- 弹出新的控制台窗口

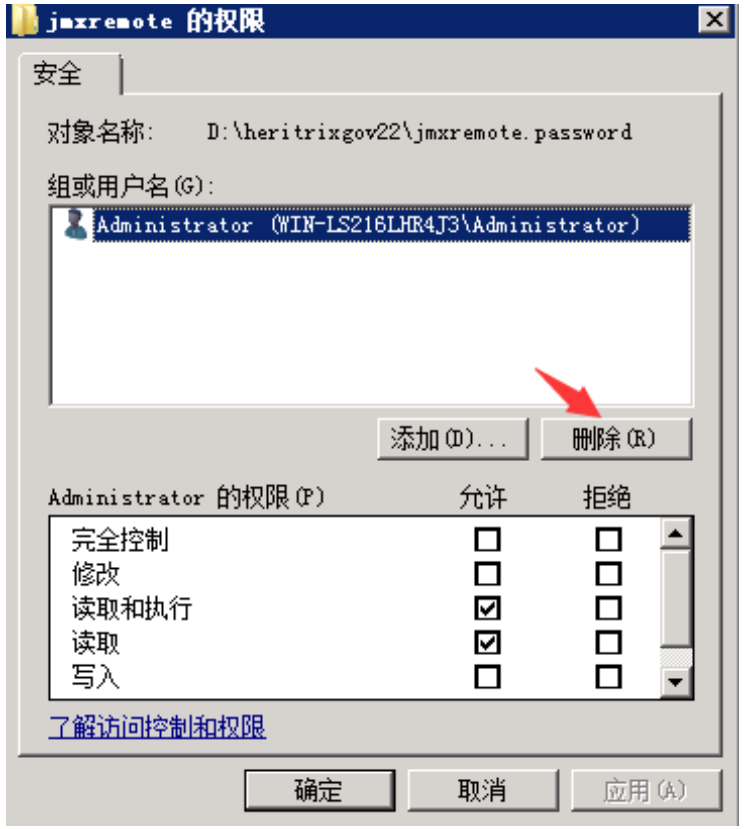
```
错误: 必须限制口令文件读取访问权限: \heritrixgov22\jmxremote.password
请按任意键继续. . .
```

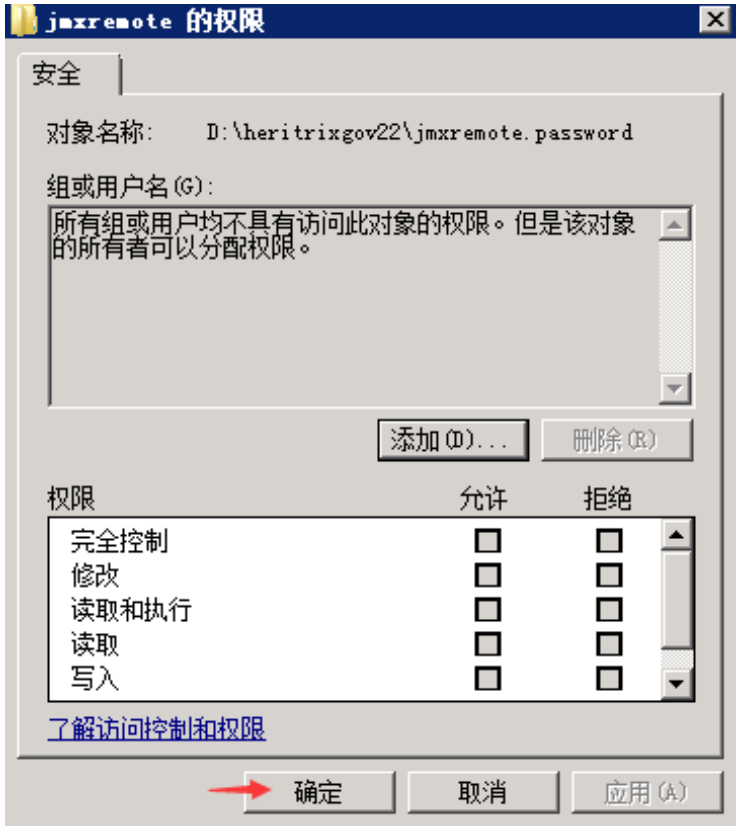
五、网页采集和网页发布软件

启动Heritrix（问题）

- 需要删除jmxremote文件的权限







五、网页采集和网页发布软件

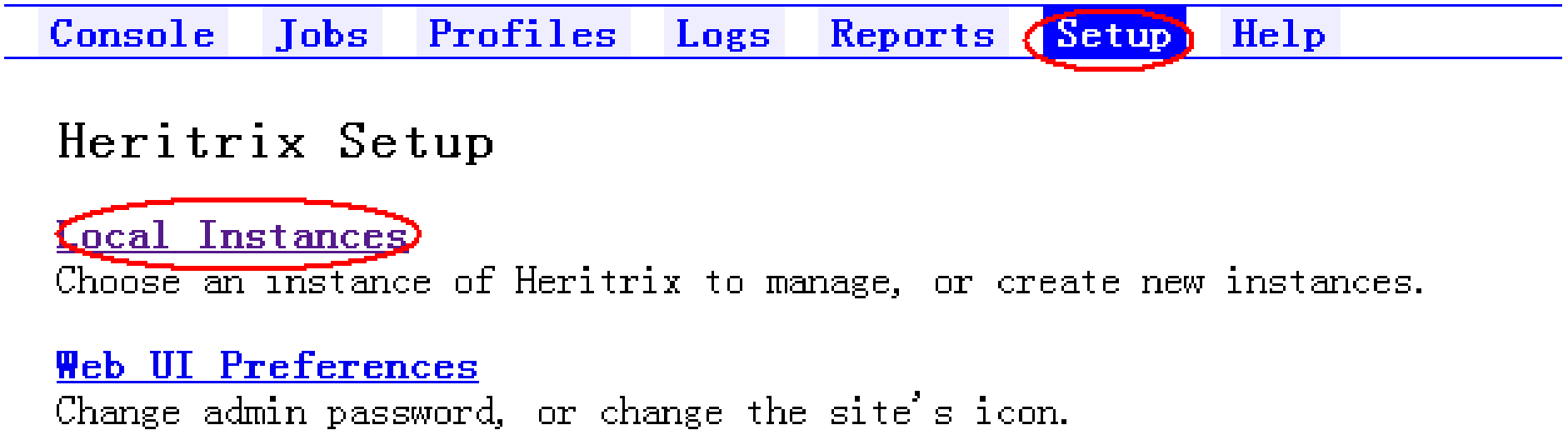
2 登陆Heritrix

- 打开浏览器，在地址栏输入Heritrix系统的访问网址http://127.0.0.1:9000，输入用户名和密码（默认均为admin），点击“Login” 。

五、网页采集和网页发布软件

3 创建任务实例

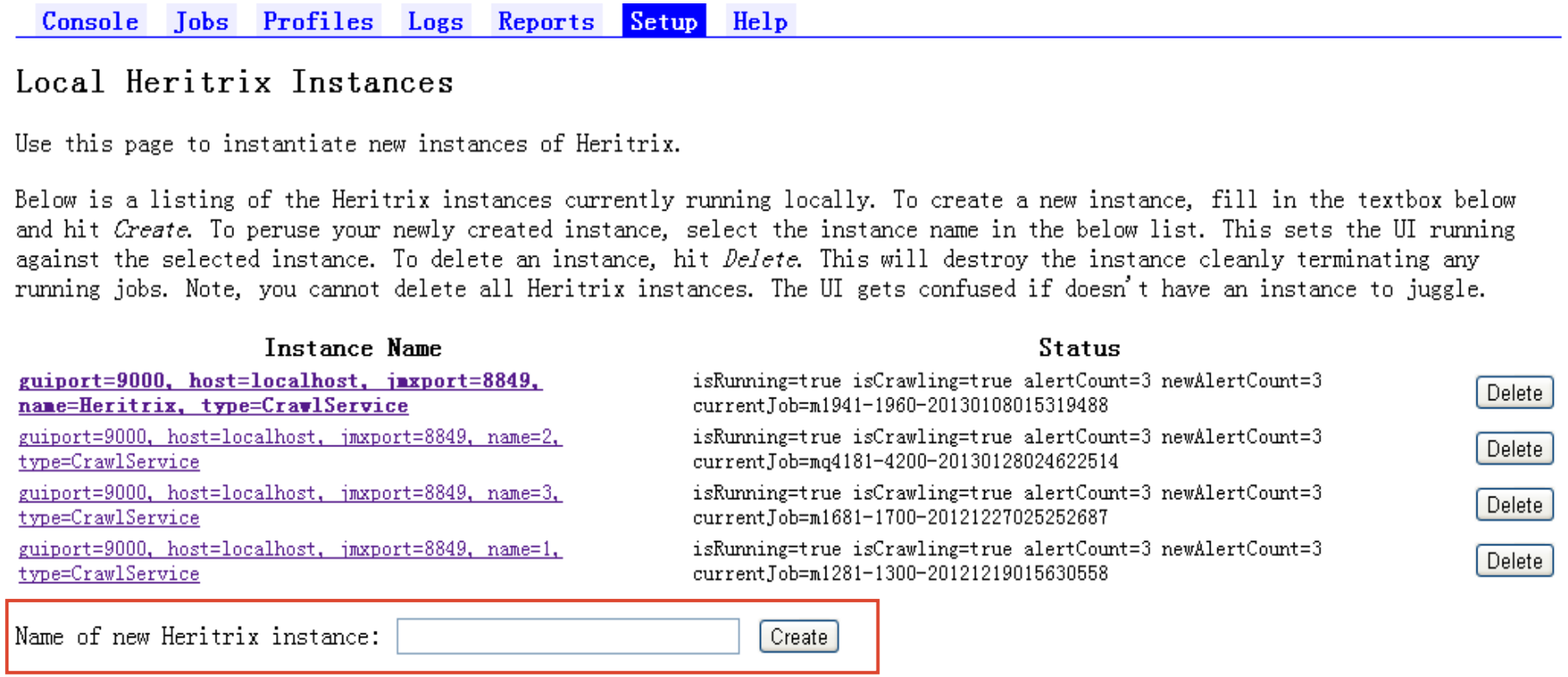
- 点击的“Setup”按钮，然后点击的“Local Instances”按钮。



五、网页采集和网页发布软件

3 创建任务实例

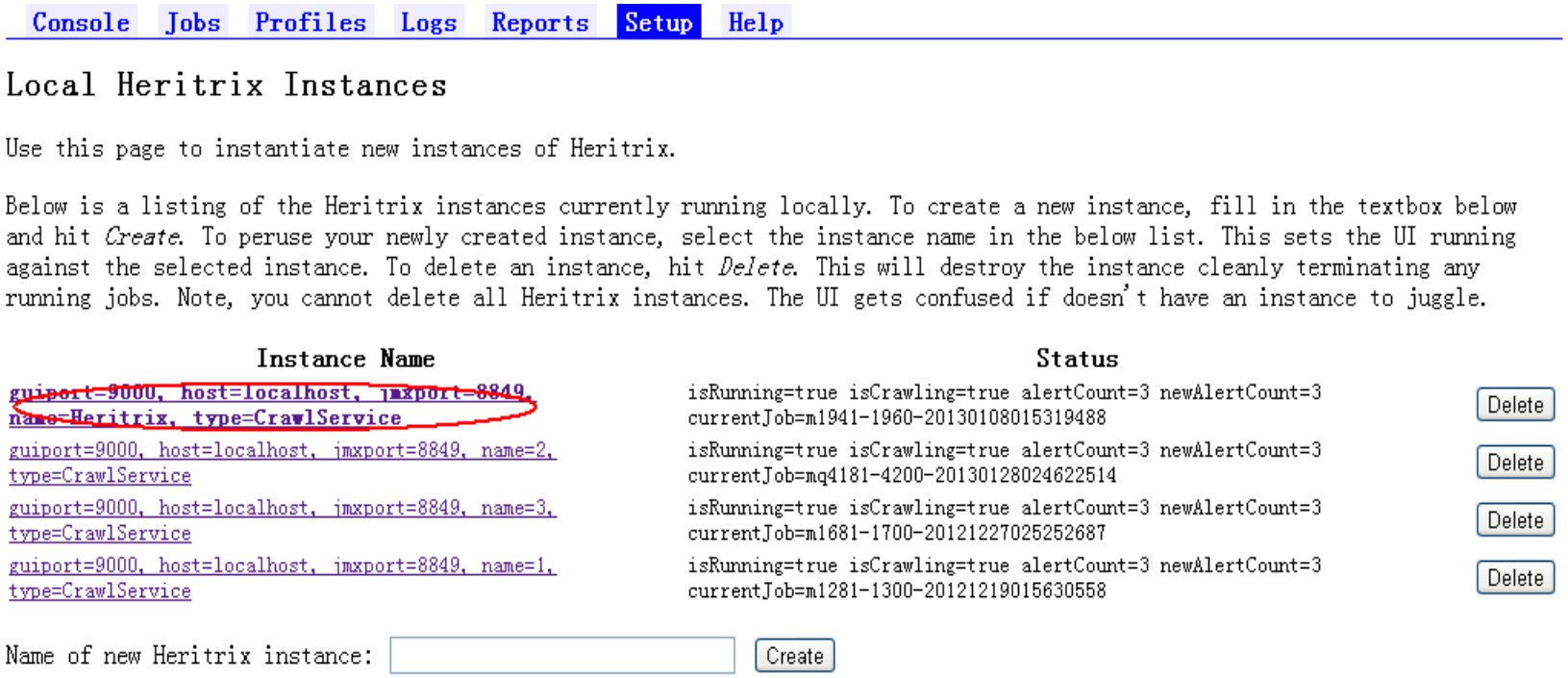
- 在 “Name of new Heritrix instance” 中输入实例名称，然后点击“Create”；一般创建4、5个实例。



五、网页采集和网页发布软件

4 创建采集任务

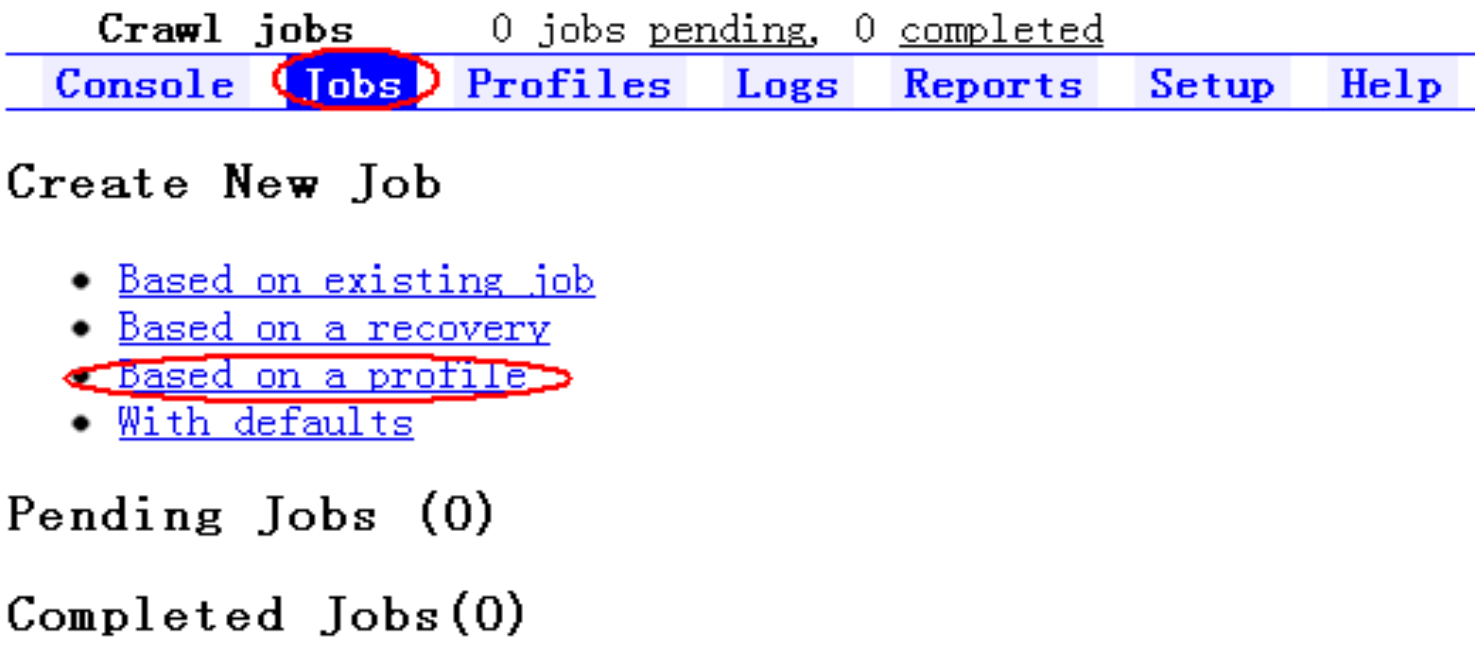
- 在Local Heritrix Instances页面中，选中一个实例的名称。



五、网页采集和网页发布软件

4 创建采集任务

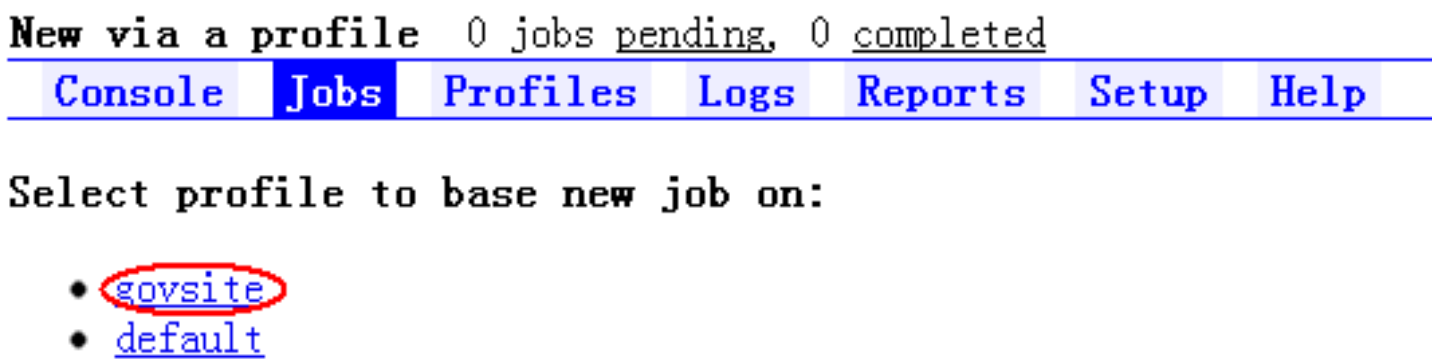
- 点击 “Jobs”按钮，在新打开的页面中点击页面中间“Create New Job”下的“Based on a profile”按钮。



五、网页采集和网页发布软件

4 创建采集任务

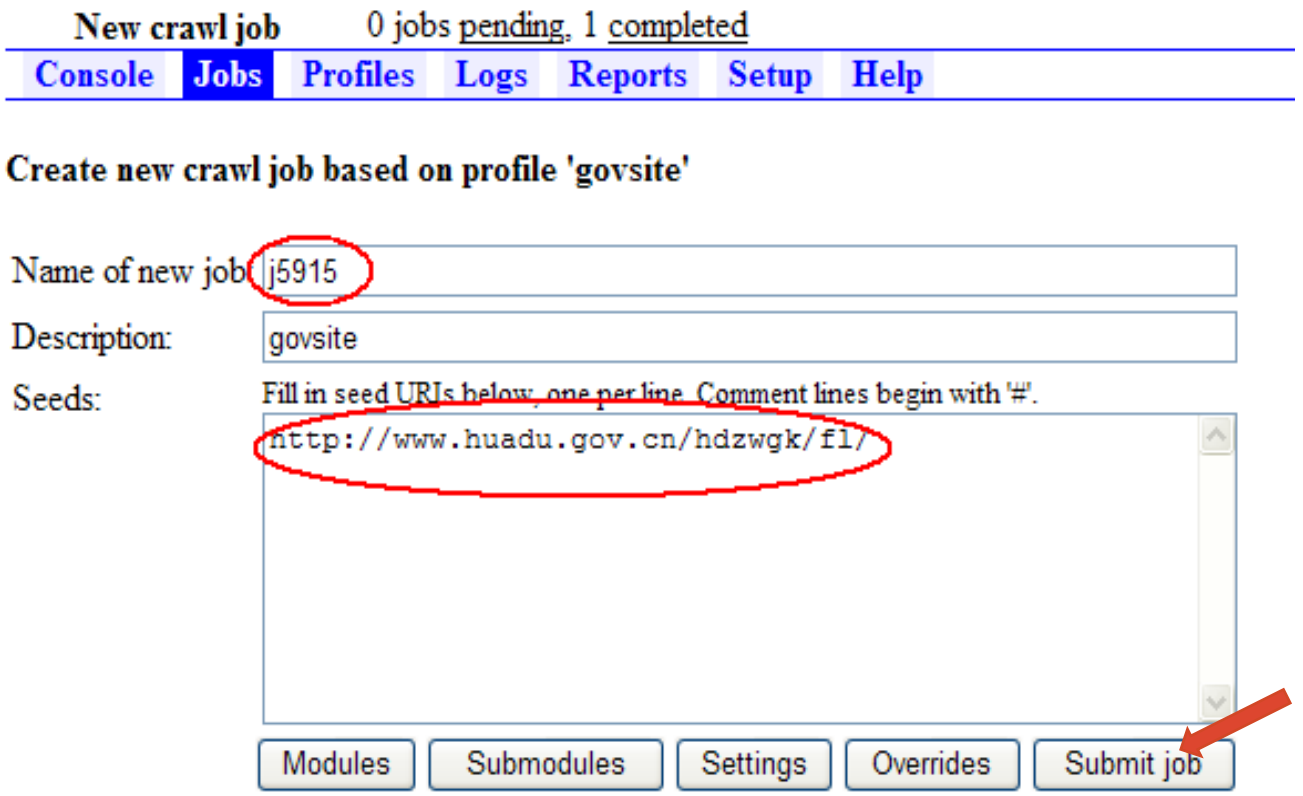
- 选择 “govsite”采集模板，并点击打开。



五、网页采集和网页发布软件

4 创建采集任务

- 在“Name of new job”输入任务名称；在“Seeds” 输入要采集的网址。

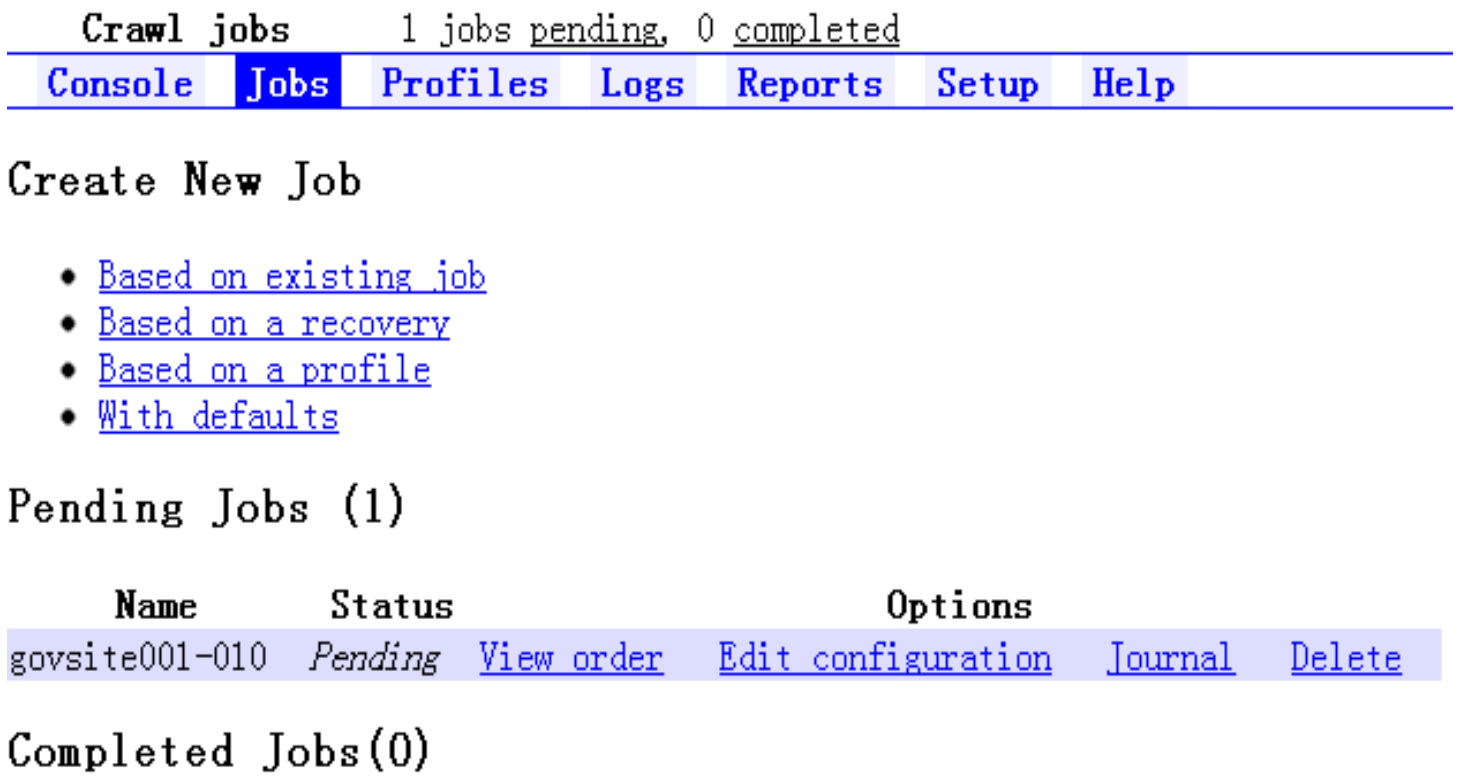


注：Seeds文本框中的URL必须以“/”结尾。

五、网页采集和网页发布软件

4 创建采集任务

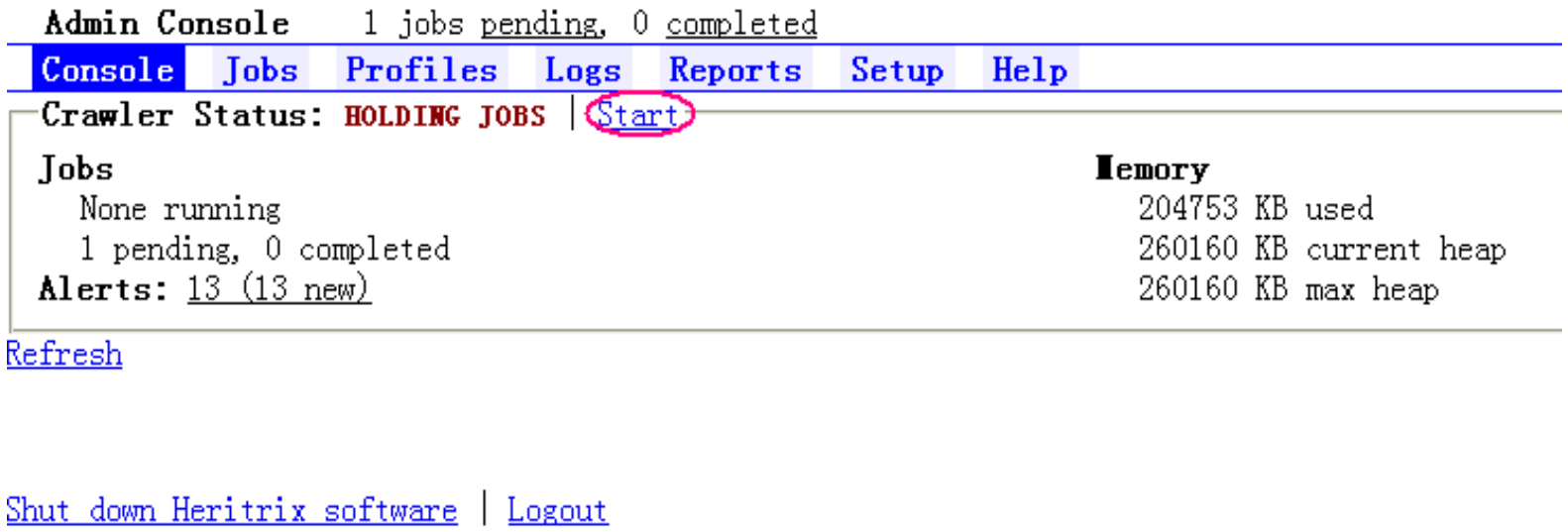
- 点击上页“Submit job”按钮，完成一个任务的部署工作。



五、网页采集和网页发布软件

4 创建采集任务

- 点击“Console”按钮，在新打开的页面中点击“Start”按钮，即可开始对网站进行采集。

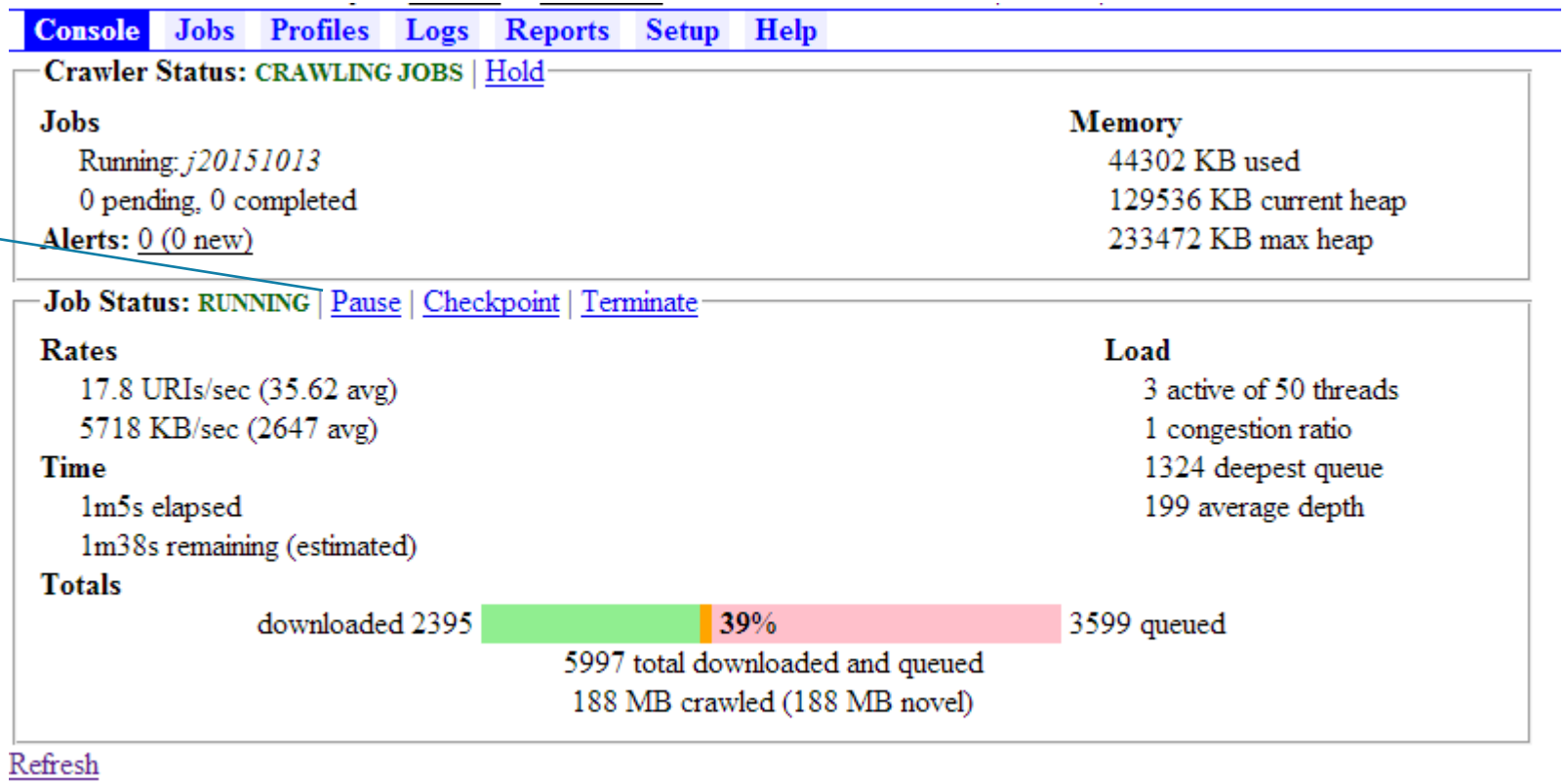


五、网页采集和网页发布软件

5 查看任务信息

- 点击“Console”按钮，可查看采集任务的运行情况。

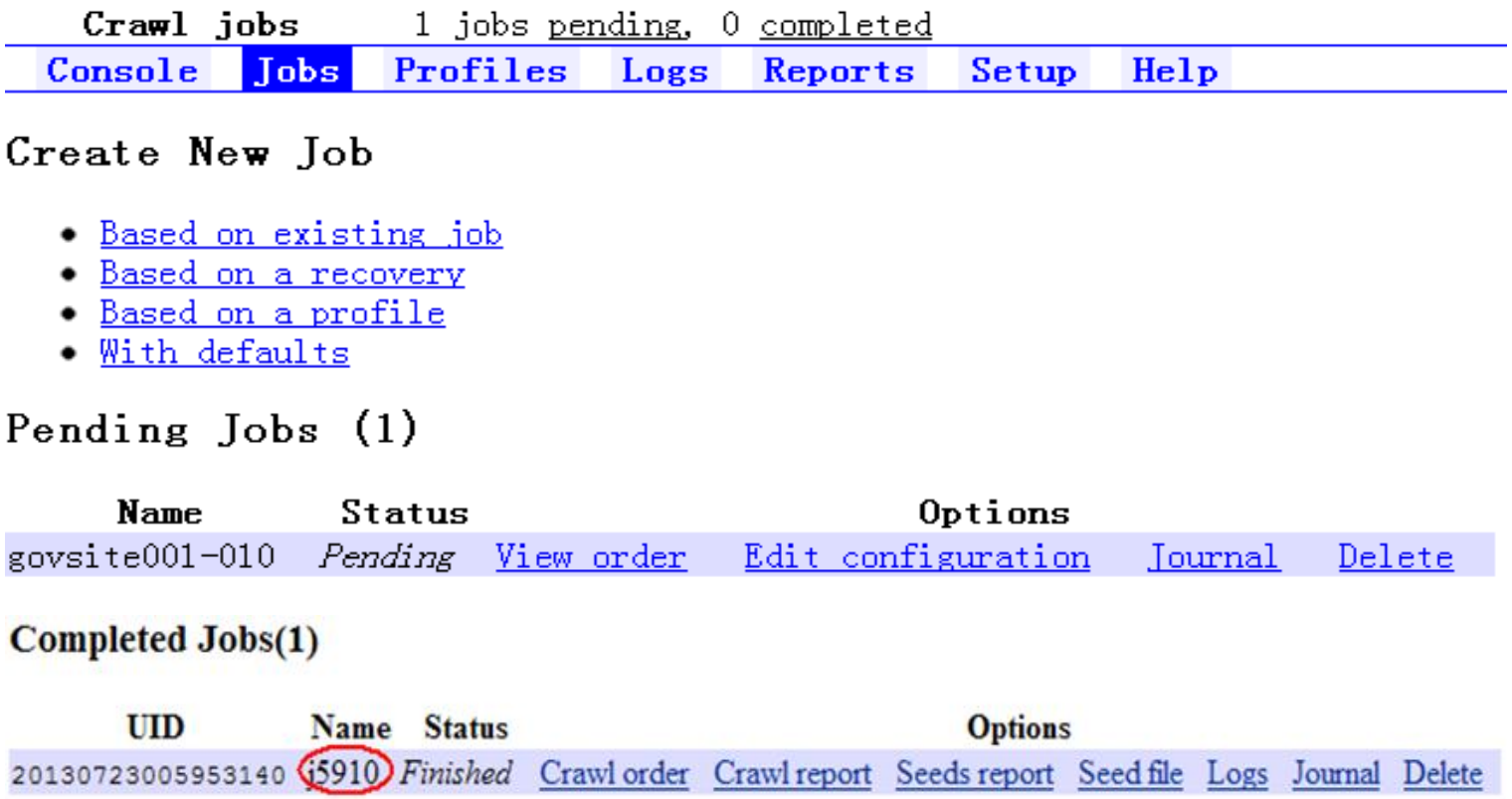
Pause：暂停采集
Checkpoint：断点采集
Terminate：终止采集



五、网页采集和网页发布软件

5 查看任务信息

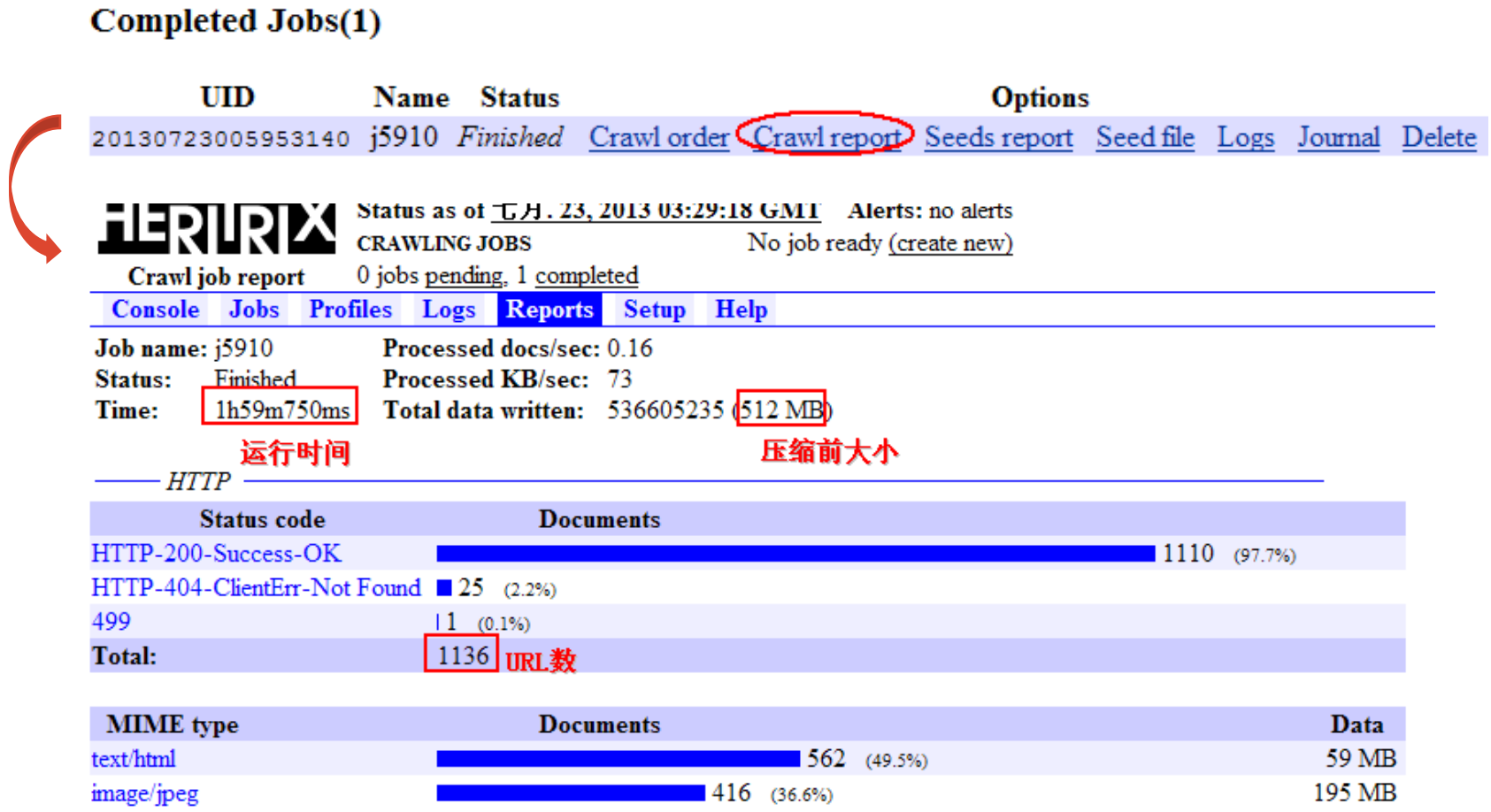
- 点击“Jobs”按钮，在打开的页面中可以查看采集任务的完成情况。



五、网页采集和网页发布软件

5 查看任务信息

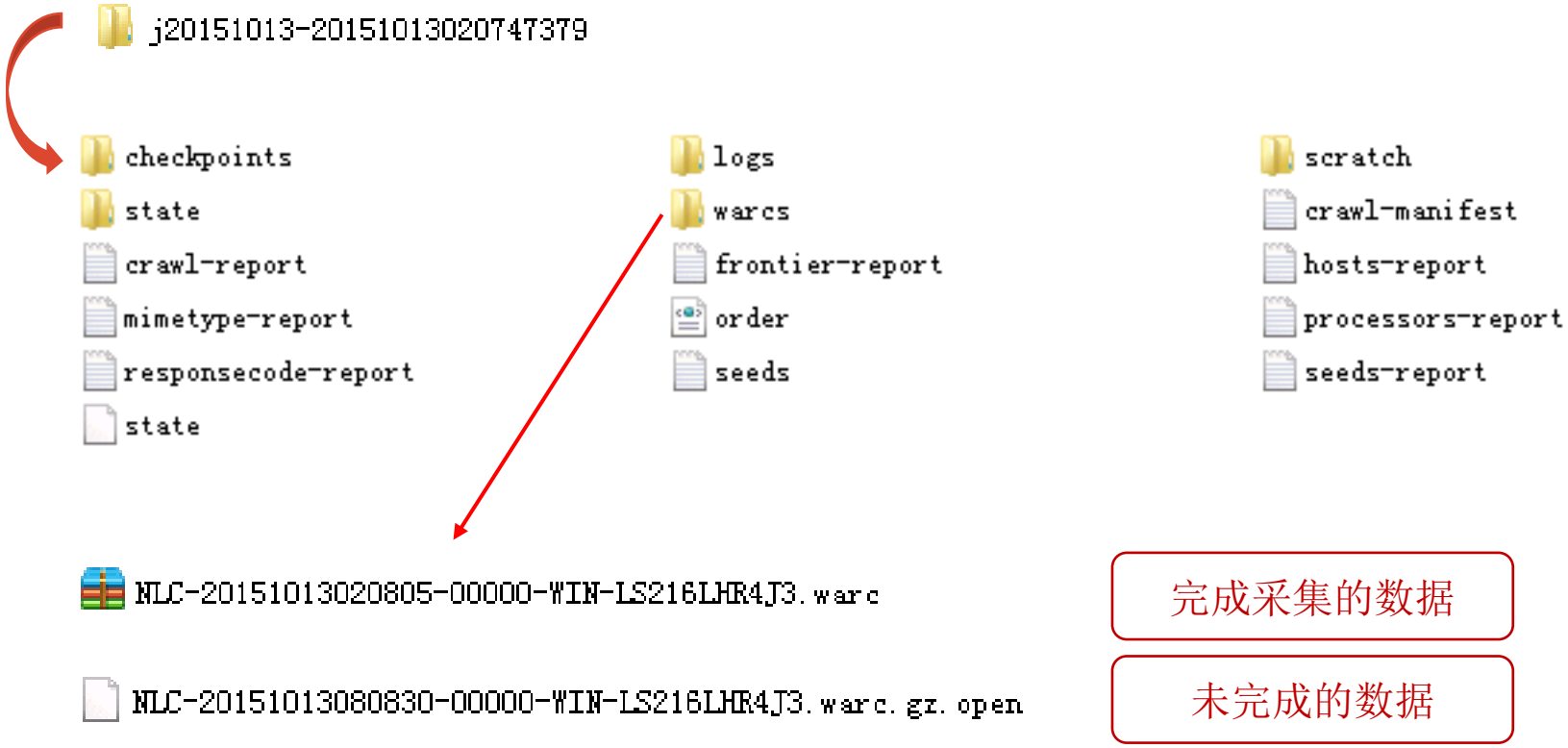
- 点击完成任务行的“Crawl report”按钮。



五、网页采集和网页发布软件

6 查看对象数据

- 在Heritrix的jobs目录下



五、网页采集和网页发布软件

网页发布软件

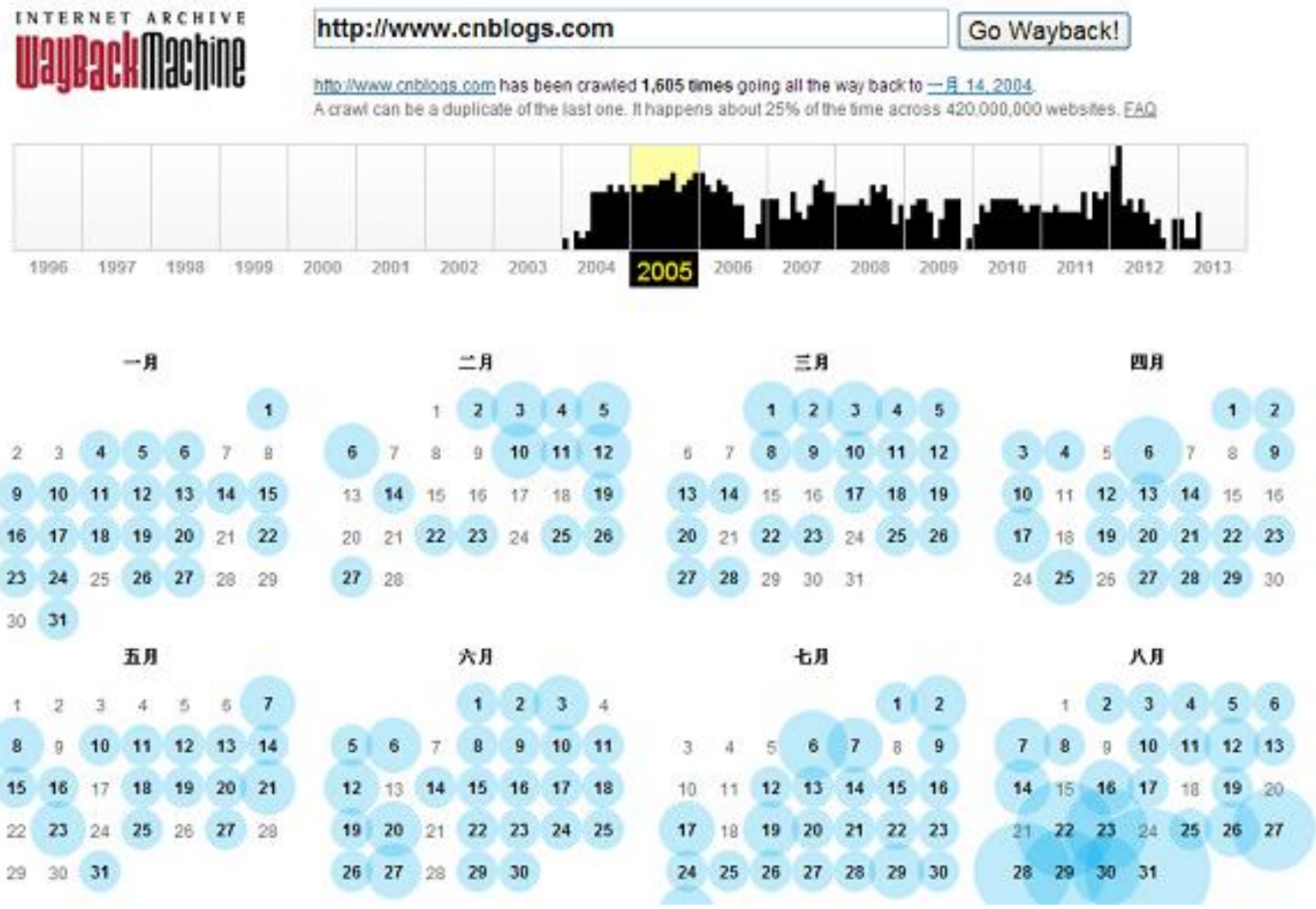
- Wayback Machine
- 网站时光倒流机
- Apache Tomcat
- 1996年开始
- 4千亿网页

Apache Tomcat

五、网页采集和网页发布软件

网页发布软件

https://archive.org/web/



五、网页采集和网页发布软件

软件准备

- 安装Apache Tomcat

  将Tomcat压缩包文件解压缩

- 配置Java和Tomcat的环境变量（可参考通用配置方法）

  在Tomcat的conf目录下，打开server.xml文件，可修改默认端口号。

```
<Connector port="8080" protocol="HTTP/1.1"
           connectionTimeout="20000"
           redirectPort="8443" />
```

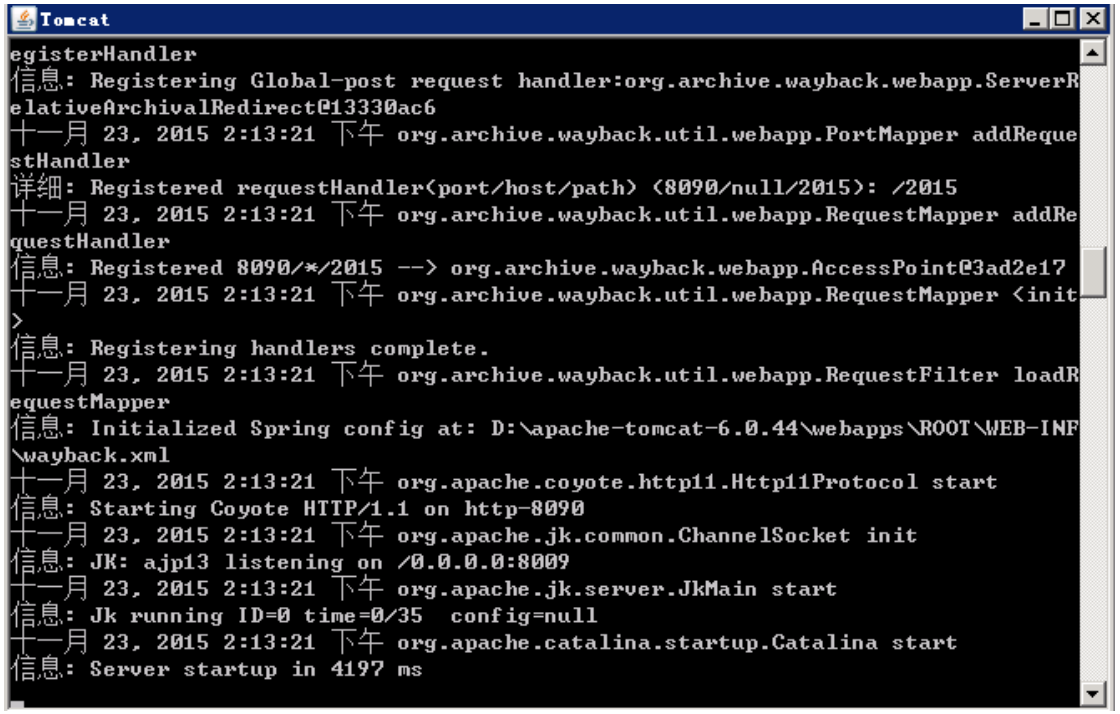
```
<Connector port="8090" protocol="HTTP/1.1"
           connectionTimeout="20000"
           redirectPort="8443" />
```

http://tomcat.apache.org/

五、网页采集和网页发布软件

1 安装 Wayback

- 将Wayback压缩包文件拷贝至Tomcat的webapps目录下。
- 打开Tomcat的bin目录，运行startup程序。

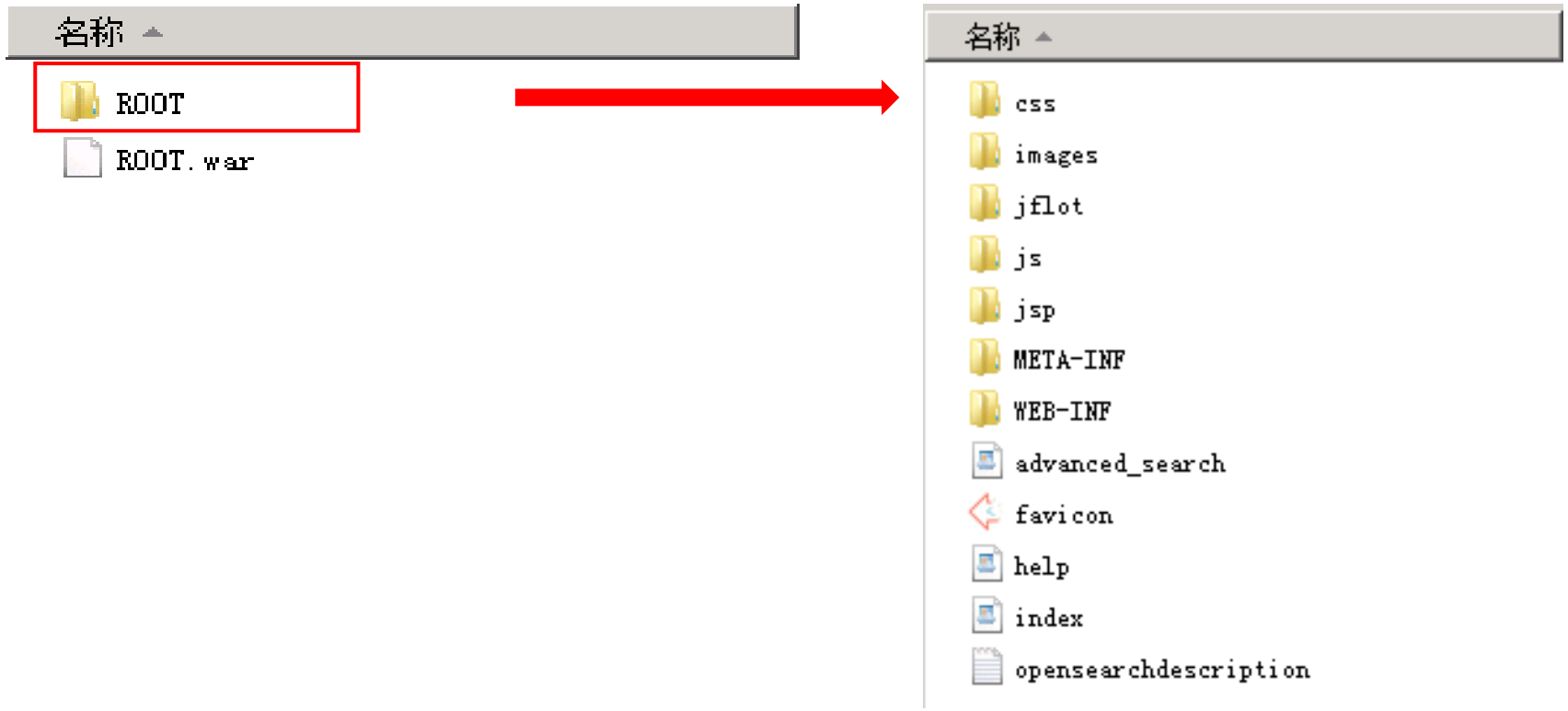


如果运行框一闪而过，说明环境变量配置错误，需要检查配置情况。

五、网页采集和网页发布软件

1 安装 Wayback

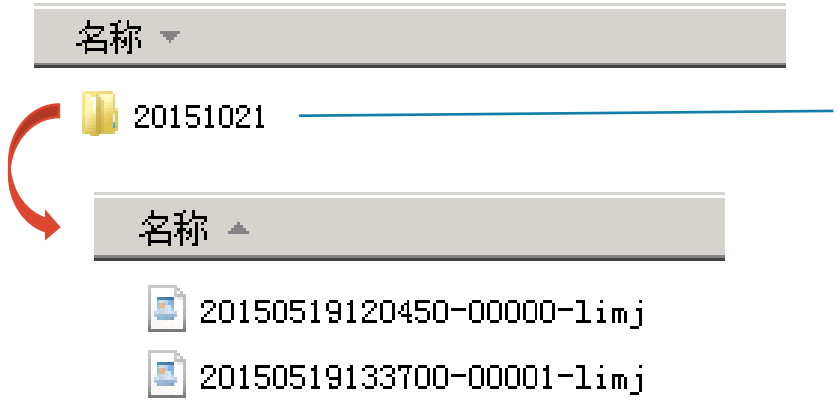
- Wayback压缩包文件会自动解压，生成一个ROOT文件夹，即完成安装。



五、网页采集和网页发布软件

2 拷贝文件

- 将采集好的WARC文件拷贝至服务器的默认位置，Tomcat安装盘符根目录下的/tmp/openwayback。



WARC文件夹的名称需要在wayback.xml中进行配置

五、网页采集和网页发布软件

3 配置文件参数

Tomcat的webapps/ROOT/WEB-INF目录下

wayback.xml

DBDCollection.xml

五、网页采集和网页发布软件

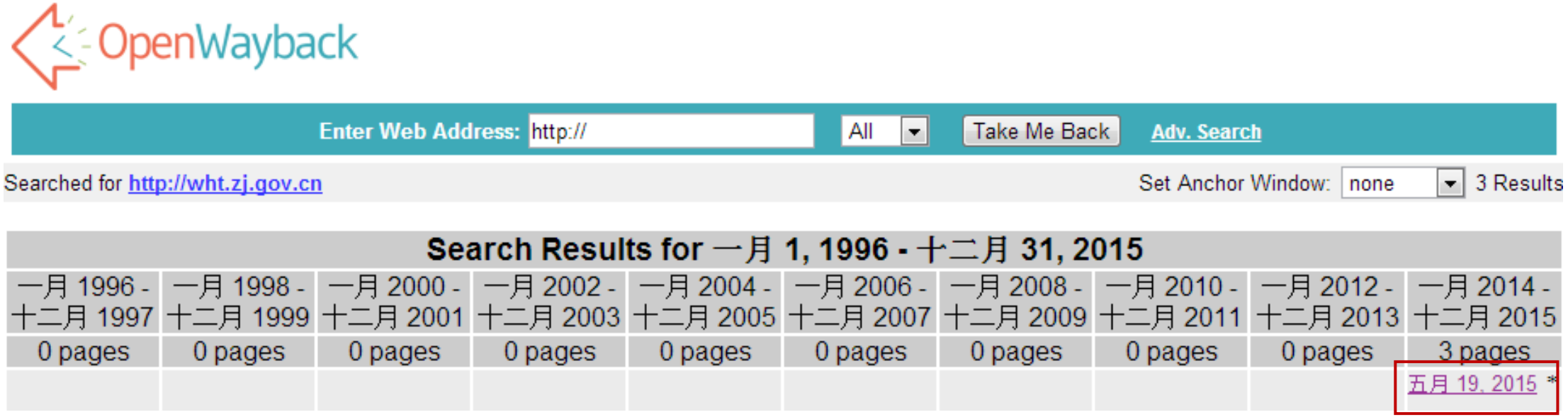
4 重启Tomcat

- 打开Tomcat的bin目录，运行shutdown关闭程序。

- 运行startup启动程序。

五、网页采集和网页发布软件

5 访问页面



谢谢！

联建网事典藏项目QQ 群：

365776635