

Web Curator Tool Lesson Plan

IIPC 2019 - WCT Workshop - Hands-on Session

[Introduction](#)

[Core Concepts](#)

[Lessons](#)

[Tutorial A - Basic Crawling](#)

[Setup Harvest Authorisation](#)

[Configure default profile](#)

[Setup Target](#)

[Monitor the Target Instance](#)

[Quality review](#)

[Tutorial B - Crawl Scoping](#)

[Setup Rejection Reasons](#)

[Setup Target](#)

[Monitor Target Instance](#)

[Adjust the Target](#)

[Compare and review the crawl](#)

[Tutorial C - Advanced Harvest Authorizations](#)

[Create a Permission Request Template](#)

[Create Harvest Authorisation](#)

[Send Permission Request](#)

[Revisiting the Harvest Authorisation](#)

[Attempting a crawl](#)

[Tutorial D - Advanced Crawling and Quality Review](#)

[Importing Profiles](#)

[Groups](#)

[Setup Target](#)

[Heritrix 3 Scripting Console](#)

[Quality review](#)

[Tutorial E - User Management](#)

[Creating the role](#)

[Create the user](#)

[Tutorial F - Reporting](#)

Introduction

This Web Curator Tool (WCT) lesson plan is comprised of a series of tutorials created for use within a workshop setting. They are designed to teach a basic understanding of the functionality and capability of WCT, without exposing any background system administration tasks.

To complete the tutorials in this lesson, you will need access to a working setup of WCT, that includes OpenWayback and Heritrix 3. Participants of an officially-run WCT workshop will be provided with a Virtual Machine environment containing these as well as any setup instructions.

The URLs for two live websites are required for completing the crawling components in this lesson plan.

- One should be a common community website. Unless otherwise told, please use <http://netpreserve.org/>
- One should be your own institution's website, that you have authority to crawl.

Please follow the profile instructions in the tutorials, so as not to overload these websites with too much crawler traffic.

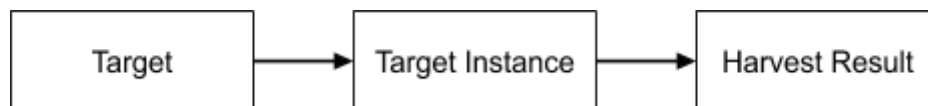
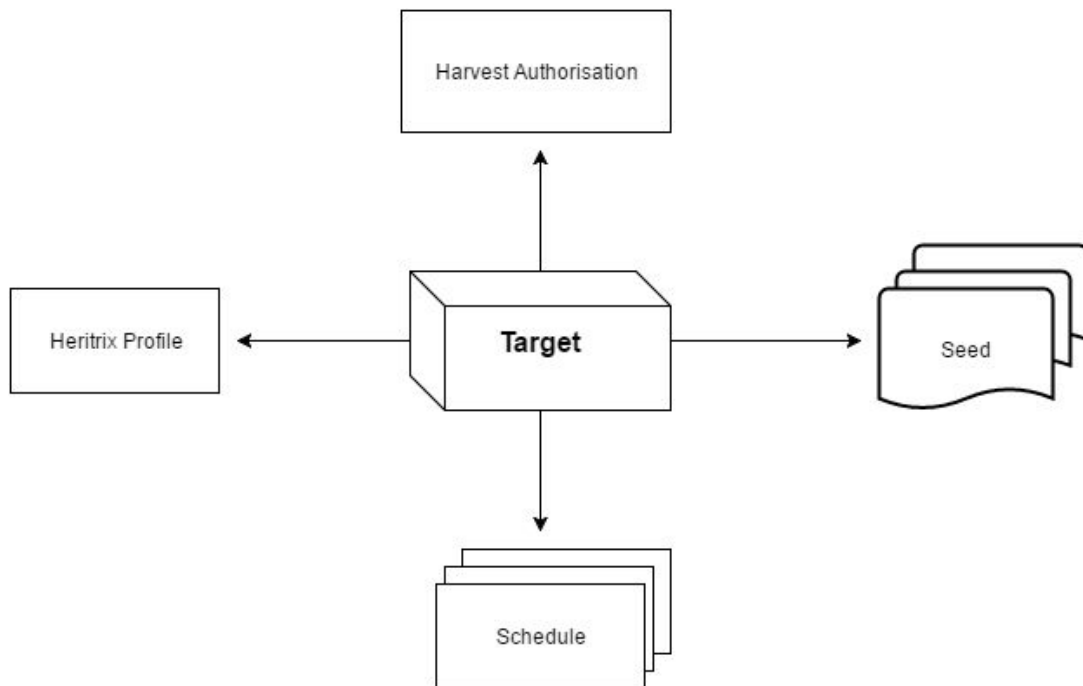
This lesson plan was written by the National Library of New Zealand and National Library of the Netherlands for the WCT Workshop held at the 2019 IIPC General Assembly, based on version 2.x of WCT.

Core Concepts

Before starting the tutorials, please familiarize yourself with the following core concepts about the WCT taxonomy/structure.

Target	The desired entity that is being harvested. Contains information relevant to the harvest including some metadata.
Schedule	When to run the harvest. A Target may have many schedules.
Profile	A profile contains the settings that control how a harvest behaves. This is the configuration that Heritrix uses during the crawl.
Harvest Authorisation	A record of the permission obtained from the copyright owner for the target of the harvest.
Seed	A URL to be harvested. A Target may have many seeds.
Target Instance	Target Instances are the record of individual harvests that are scheduled from Targets.
Harvest Result	The files that are retrieved during a harvest.

The following diagrams shows the basic overview of the WCT taxonomy for a Target and its relationships:



For further reading, please see the WCT [documentation](#).

Lessons

The following tutorials are ordered in difficulty, gradually introducing more advanced functionality of WCT. Tutorials A-D focus on crawling and to a lesser extent harvest authorisations and quality review. Tutorials E-F focus more on administration of WCT.

Tutorial A - Basic Crawling

What you will learn [authorisations, profiles, scheduling, crawl monitoring, quality review]

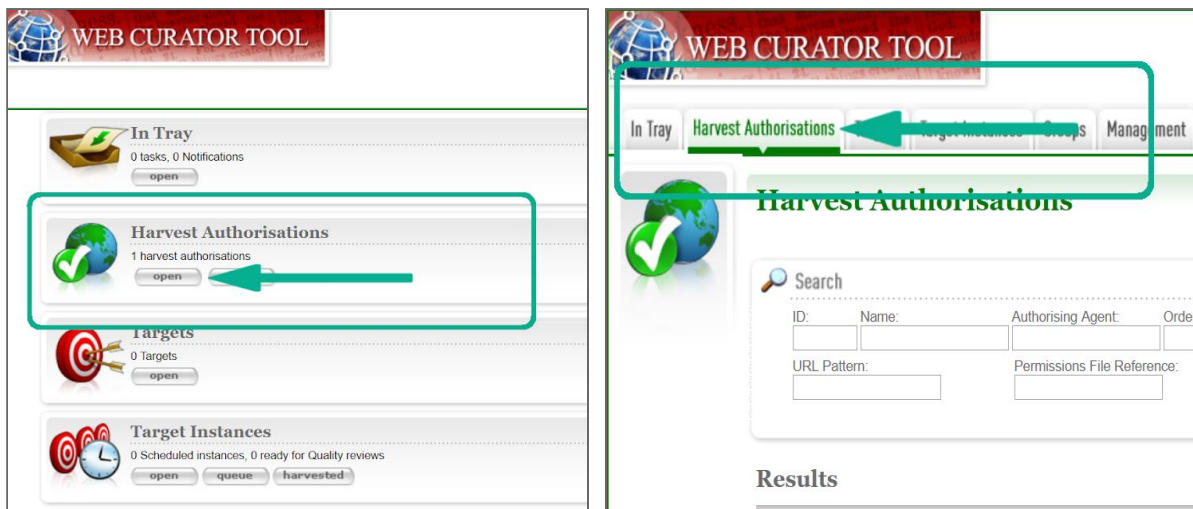
Scenario: You must crawl a website that your institution has determined requires permission to harvest.

Setup Harvest Authorisation

A Harvest Authorisation is a record of requested permission from copyright holders (e.g. website owner) to harvest a website. You must consider copyright law when you harvest material, and also when you preserve it and when you make it accessible to users.

For more information on Harvest Authorisations in WCT, see the [documentation](#).

1. Navigate to the Harvest Authorisation tab in WCT



2. Create a new Harvest Authorisation

WEB CURATOR TOOL

Home | Queue

In Tray **Harvest Authorisations** Targets Target Instances Groups Management

Harvest Authorisations

Search

ID: Name: Authorising Agent: Order No: Agency: workshop Sort Order: Name (ascending) Show Disabled:

URL Pattern: Dismissions File Reference: Dismissions Status: Pending Requested Approved Rejected

search reset

Results

Id	Created	Name	Auth Agent	Order No	Status	Action
65536	07/05/19	Blanket Authorisation	Dummy Authorising Agency		Approved	

create new

3. In the General tab enter the following detail

Title: netpreserv.org

Description: IIPC Website

WEB CURATOR TOOL

Home | Queue | Help

In Tray **Harvest Authorisations** Targets Target Instances Groups Management

Harvest Authorisations

General URL Patterns Authorising Agencies Permissions

Agency: workshop

Title: netpreserv.org

Description: IIPC website

Order No:

Published:

Enabled:

4. Navigate to the URL Patterns tab. You have determined that there are two different URLs that can be used to resolve the site. Enter the following text and click add.

*New URL Pattern: http://netpreserve.org/**

*New URL Pattern: https://netpreserve.org/**

WEB CURATOR TOOL

Home | Queue | Harvested | Help | Logout
User wct is logged in.

In Tray **Harvest Authorisations** Targets Target Instances Groups Management

Harvest Authorisations

netpreserve.org

General URL Patterns Authorising Agencies Permissions

New URL Pattern: https://netpreserve.org/* add

URL Pattern	Action
http://netpreserve.org/*	

save cancel

5. Navigate to the Authorising Agencies tab, and click Create New.

The screenshot shows the 'WEB CURATOR TOOL' interface. At the top, there are navigation links: Home | Queue | Harvested | Help | Logout, with a note 'User wct is logged in.'. Below this is a menu bar with 'In Tray', 'Harvest Authorisations', 'Targets', 'Target Instances', 'Groups', and 'Management'. The 'Harvest Authorisations' section is active, showing a globe icon with a checkmark and the title 'Harvest Authorisations' for 'netpreserve.org'. There are two tabs: 'Authorising Agencies' (selected) and 'Permissions'. A 'create new' button is highlighted with a green arrow. Below the tabs is a table with columns 'Authorising Agency', 'Contact', and 'Action'. At the bottom are 'save' and 'cancel' buttons.

6. The Authorising Agency is the organisation or individual who holds the copyright for the website you are harvesting. This record holds the primary contact for the Authorising Agency. Enter the following detail, and click Save.

Title: IIPC

Description: International Internet Preservation Consortium

Contact: Jane Doe

Email: jane.doe@test.email.address

The screenshot shows the 'WEB CURATOR TOOL' interface. At the top, there are navigation links: Home | Queue | Harvested | Help | Logout, with a note 'User wct is logged in.'. Below this is a menu bar with 'In Tray', 'Harvest Authorisations', 'Targets', 'Target Instances', 'Groups', and 'Management'. The 'Harvest Authorisations' section is active, showing a globe icon with a checkmark and the title 'Harvest Authorisations' for 'netpreserve.org'. There are two tabs: 'Authorising Agencies' (selected) and 'Permissions'. The form fields are filled with the following information: Name: IIPC, Description: International Internet Preservation Consortium, Contact: Jane Doe, Phone: (empty), Email: jane.doe@test.email.address, Address: (empty). The 'save' button is highlighted with a green arrow. At the bottom are 'save' and 'cancel' buttons.

7. Navigate to the Permissions tab, and click Create New.

The screenshot shows the 'WEB CURATOR TOOL' interface. The top navigation bar includes 'Home | Queue | Harvested | Help | Logout' and 'User wct is logged in.'. Below the navigation bar, there are tabs for 'In Tray', 'Harvest Authorisations', 'Targets', 'Target Instances', 'Groups', and 'Management'. The 'Harvest Authorisations' tab is active, and the page title is 'Harvest Authorisations' with the URL 'netpreserve.org'. The 'Permissions' tab is selected, and a red 'create new' button is highlighted with a red arrow. Below the button is a table with columns: 'Status', 'Date Requested', 'Authorising Agent', 'From', 'To', 'URL Patterns', and 'Action'. At the bottom of the form area are 'save' and 'cancel' buttons.

8. A Permission is a record of permissions requested from the copyright holder for harvested particular URLs. Enter the following detail, and click Save.

Dates: <the current date> to <blank>

Status: Approved

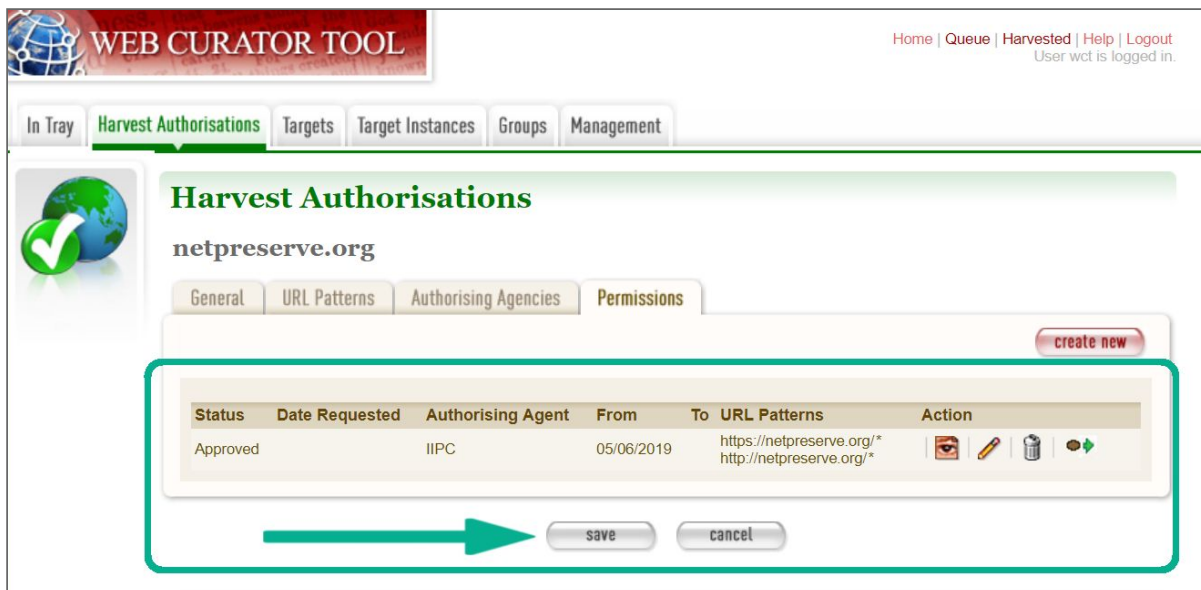
Auth. Agency Response: Hi Institution, we give permission for you to harvest our website.

Special Restrictions: Please do not crawl our website during the weekends.

Urls: <tick both>

The screenshot shows the 'WEB CURATOR TOOL' interface with the 'Harvest Authorisations' page. The 'Permissions' tab is active, and the form fields are filled with the following information: 'Authorising Agent' is 'IIPC', 'Dates' is '05/06/2019' to a blank field, 'Status' is 'Approved', 'Auth. Agency Response' is 'Hi Institution, we give permission for you to harvest our website.', 'Special Restrictions' is 'Please do not crawl our website during the weekends.', 'Copyright Statement' is empty, 'Copyright URL' is empty, 'Access Status' is 'Open (unrestricted) access', 'Open Access Date' is empty, 'Quick Pick' is unchecked, 'Display Name' is empty, 'Urls' are 'http://netpreserve.org/*' and 'https://netpreserve.org/*' (both checked), 'File Reference' is empty, and 'Assign Approval Task' is 'No'. The 'Exclusions' section at the bottom is empty, with the message 'No exclusions have been defined.'.

9. Save the new Harvest Authorisation.



WEB CURATOR TOOL

Home | Queue | Harvested | Help | Logout
User wct is logged in.

In Tray **Harvest Authorisations** Targets Target Instances Groups Management

Harvest Authorisations

netpreserve.org

General URL Patterns Authorising Agencies **Permissions**

create new

Status	Date Requested	Authorising Agent	From	To	URL Patterns	Action
Approved		IIPC	05/06/2019		https://netpreserve.org/* http://netpreserve.org/*	   

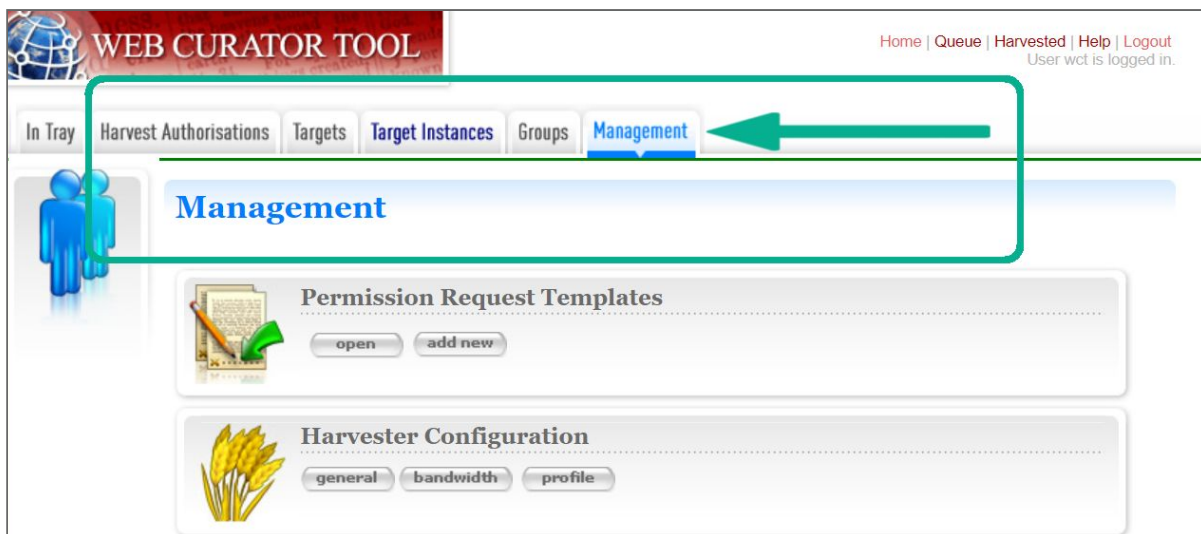
save cancel

Configure default profile

A Profile contains the settings used to control the harvest of a website. The settings for WCT profiles are based on the Heritrix crawler. There are many options available for configuring the behaviour of Heritrix, however for basic usage, WCT provides a subset of core settings through the user interface.

For more information on WCT profiles, see the [documentation](#).

1. Navigate to the Management tab in WCT





WEB CURATOR TOOL

Home | Queue | Harvested | Help | Logout
User wct is logged in.

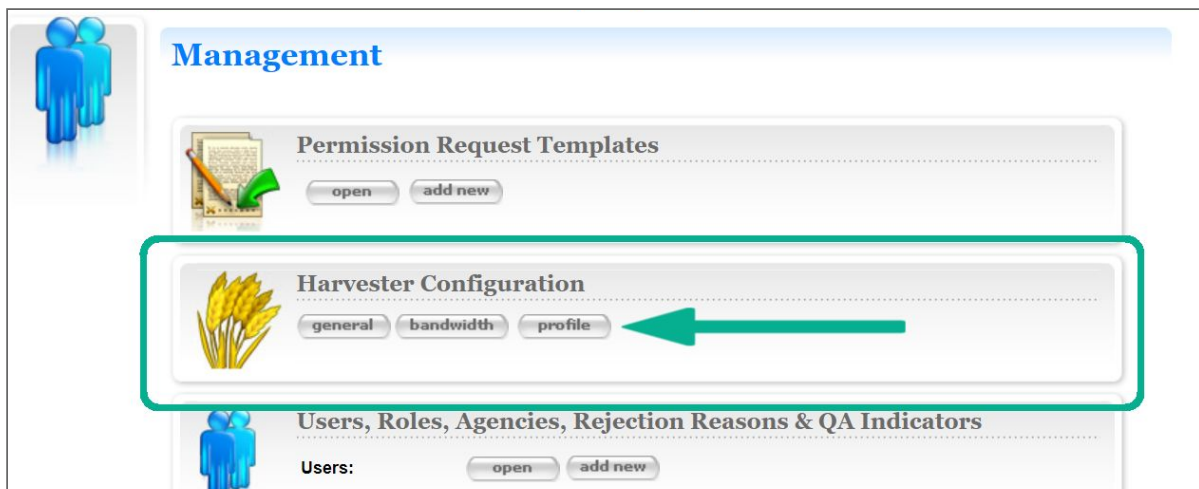
In Tray Harvest Authorisations Targets **Target Instances** Groups **Management**

Management

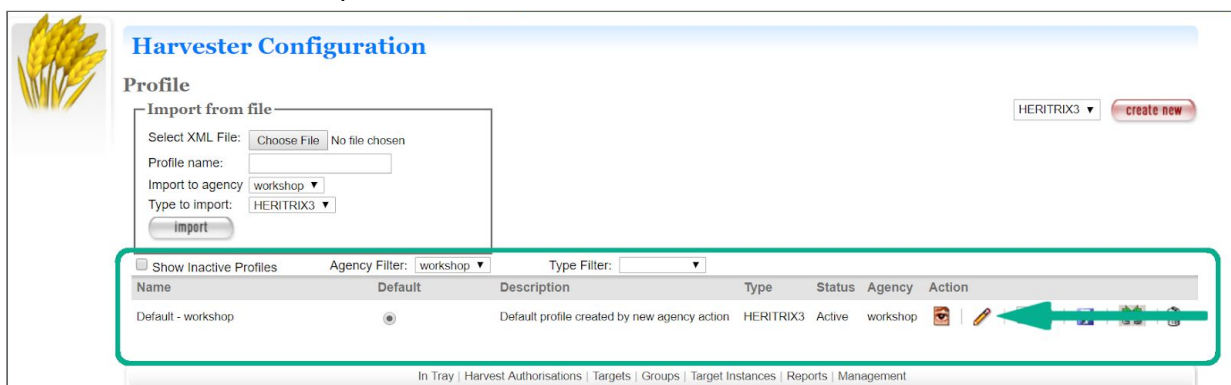
 **Permission Request Templates**
open add new

 **Harvester Configuration**
general bandwidth profile

2. Click the Profile option under Harvester Configuration.



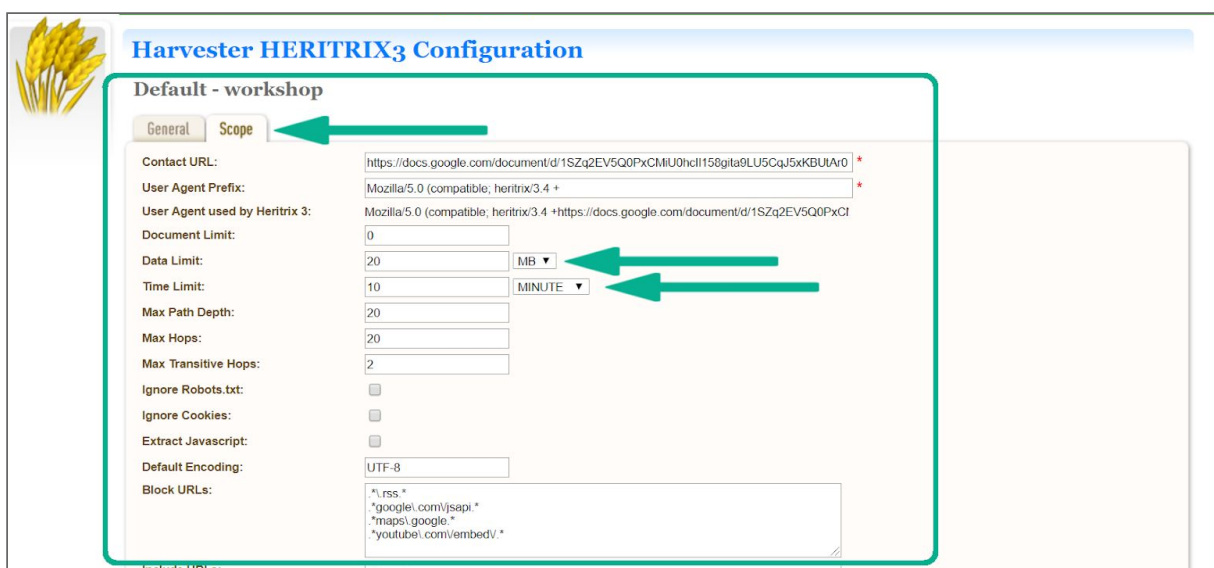
3. You should see the default profile listed - "Default - workshop". Click the edit icon next to the default profile.



4. Navigate to the Scope tab. We will set data and time limits in the profile, so we don't wait too long to continue this tutorial. Enter the following settings and click Save.

Data Limit: 20 MB

Time Limit: 10 Minute

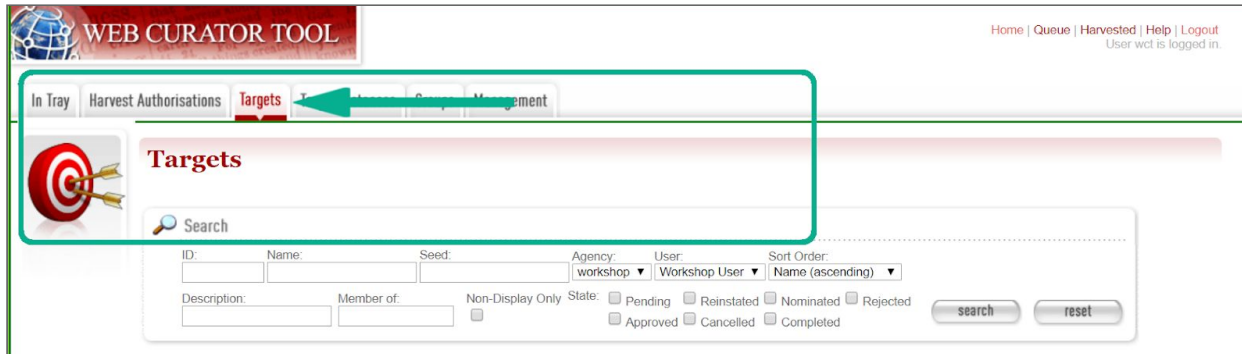


Setup Target

A Target is a record that describes the thing you intend to harvest from the Web. It contains metadata descriptions, seed URLs, how and when the thing will be crawled.

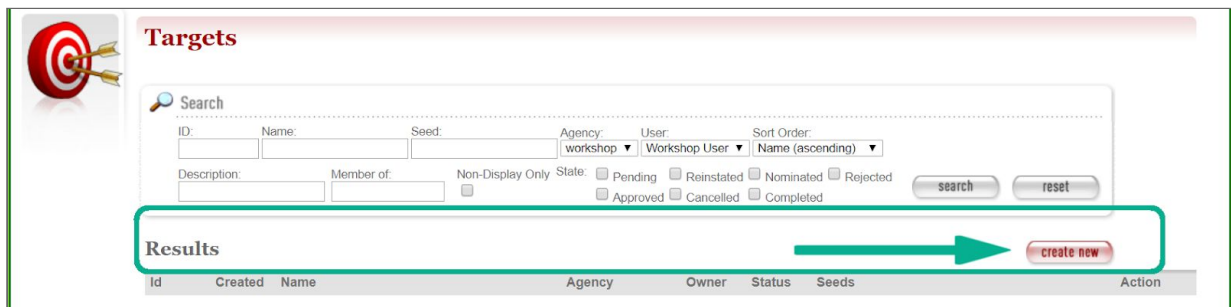
For more information on WCT Targets, see the [documentation](#).

1. Navigate to the Targets tab in WCT



The screenshot shows the 'WEB CURATOR TOOL' interface. The 'Targets' tab is highlighted in the navigation bar. A green box highlights the 'Targets' tab and the search area below it. The search area includes fields for ID, Name, Seed, Agency, User, and Sort Order, along with checkboxes for various states like Pending, Reinstated, Nominated, Rejected, Approved, Cancelled, and Completed. There are 'search' and 'reset' buttons.

2. Create a new Target



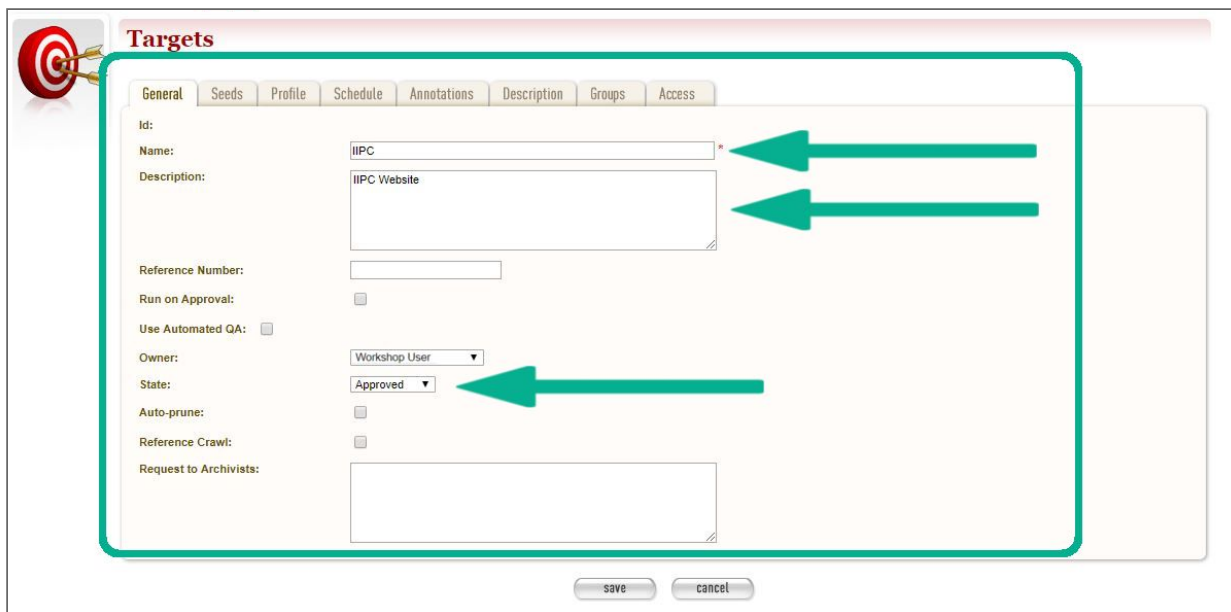
The screenshot shows the 'Targets' page with a search area and a 'Results' table. A green box highlights the 'create new' button in the 'Action' column of the table. The table has columns for Id, Created, Name, Agency, Owner, Status, Seeds, and Action.

3. In the General tab, enter the following detail

Name: IIPC

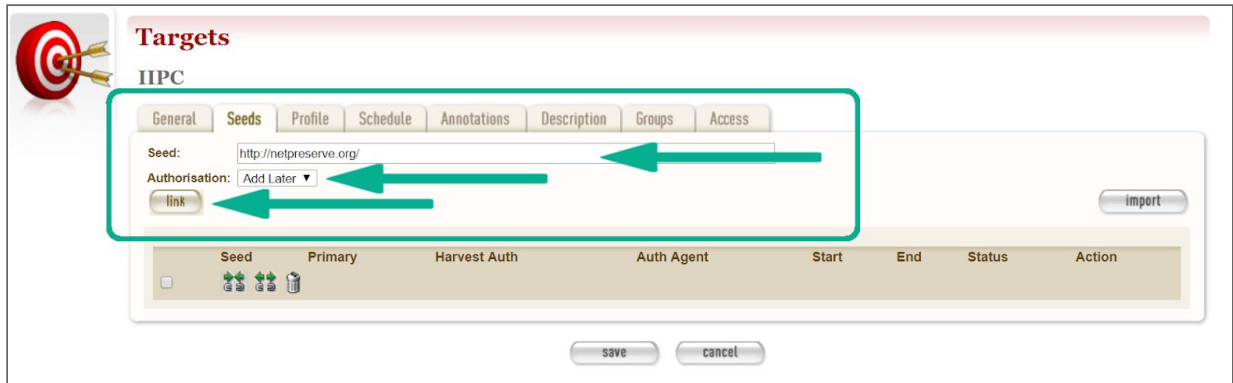
Description: IIPC Website

State: Approved

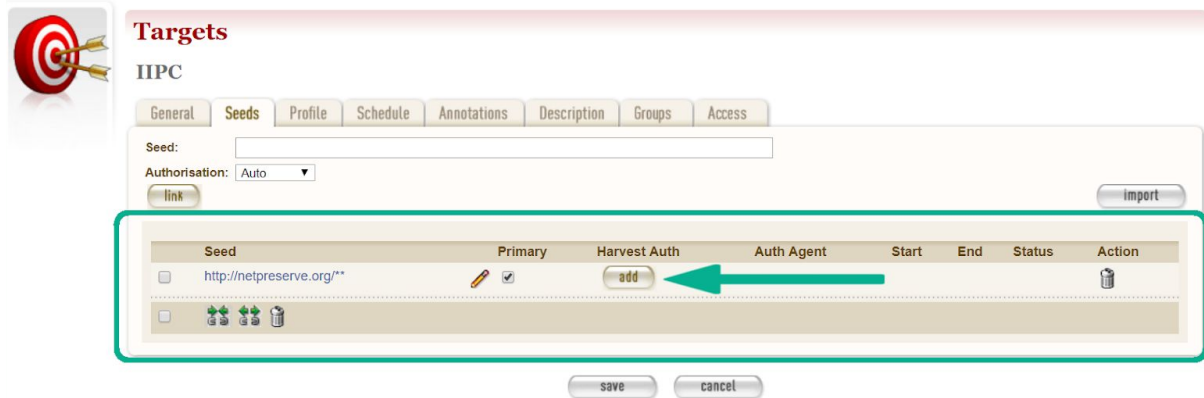


The screenshot shows the 'Targets' page with the 'General' tab selected. A green box highlights the input fields for Name, Description, and State. Green arrows point to the 'Name' field (containing 'IIPC'), the 'Description' field (containing 'IIPC Website'), and the 'State' dropdown menu (set to 'Approved'). There are 'save' and 'cancel' buttons at the bottom.

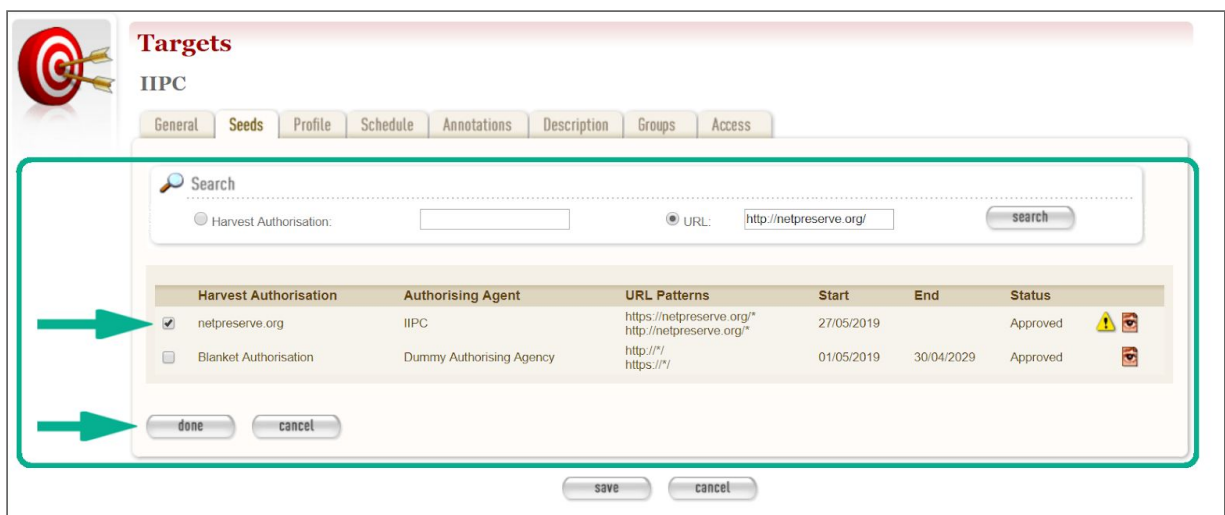
- Navigate to the Seeds tab, and enter the following detail, then click the Link button.
*Seed: http://netpreserve.org/
 Authorisation: Add Later*



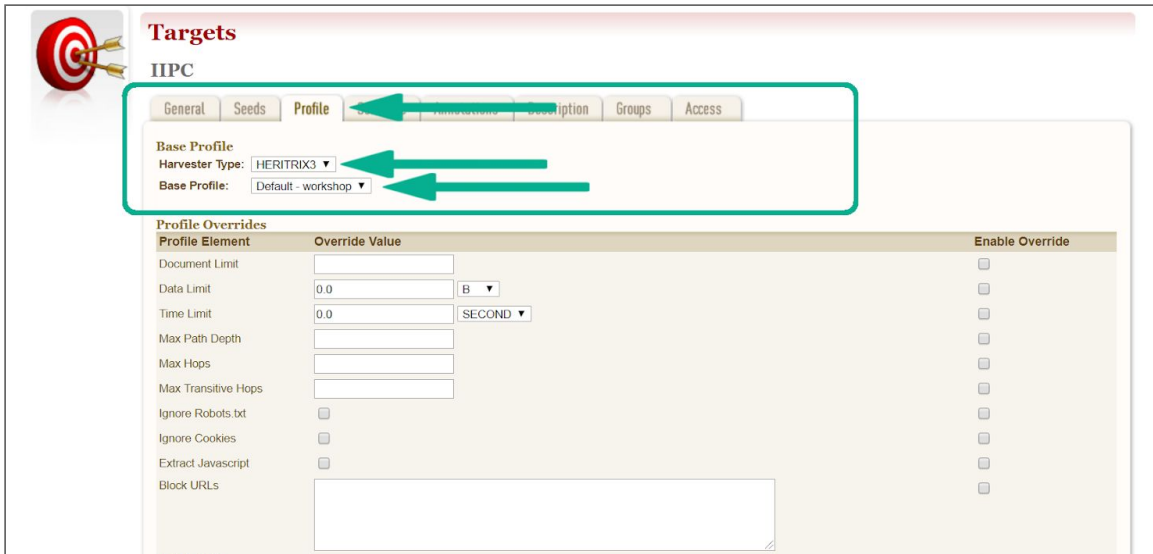
- Click the Add button to assign a Harvest Authorisation to this seed.
Each seed must be linked to an authorisation which details the permission granted to harvest content from that URL.



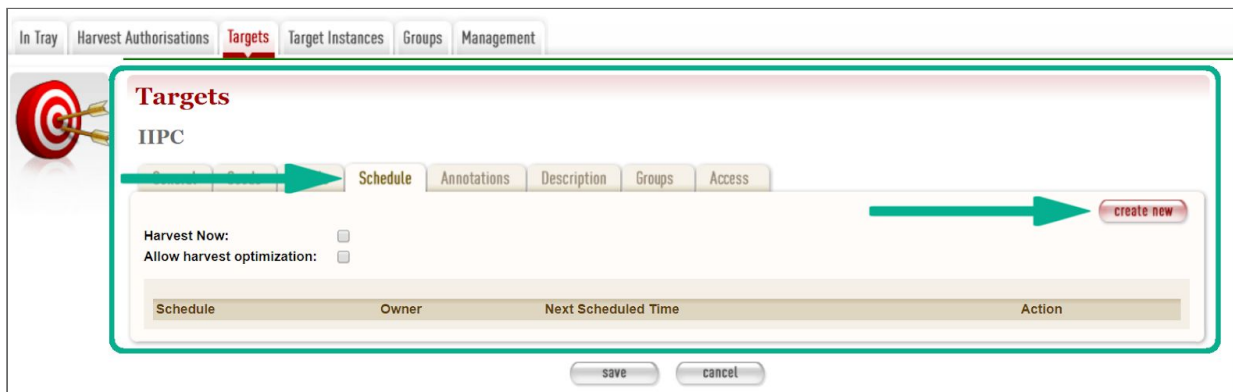
- Tick the box next to the netpreserve.org Harvest Authorisation and click Done.
Notice the yellow warning symbol, indicating that the Permission has special restrictions. Click the View icon next to it, to see the Permission details.



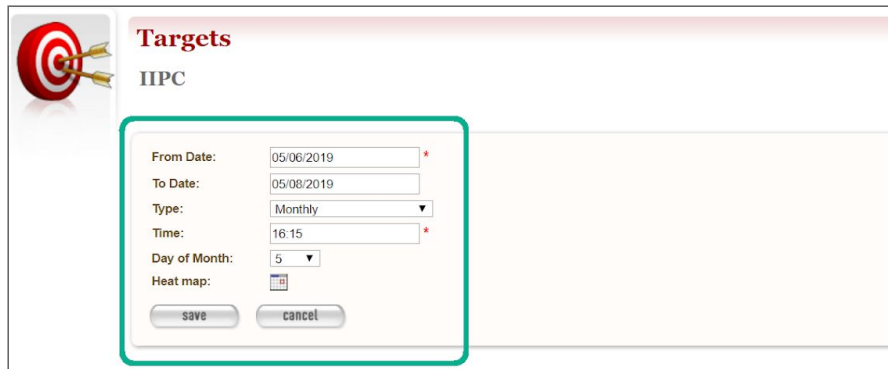
- Navigate to the Profile tab, and assign the profile to be used during the crawl. Ensure the following detail is set.
Harvester Type: HERITRIX3
Base Profile: Default - workshop



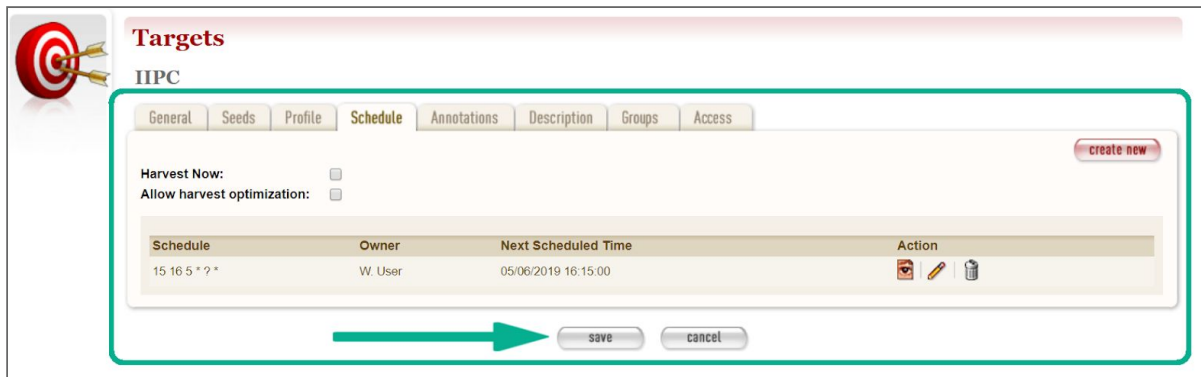
- Navigate to the Schedule tab, click Create New.



- Enter the following detail, and click Save.
From Date: <today's date>
To Date: <three months from today's date>
Type: Monthly
Time: <five minutes from current time>
Day of Month: <the current day of the month>



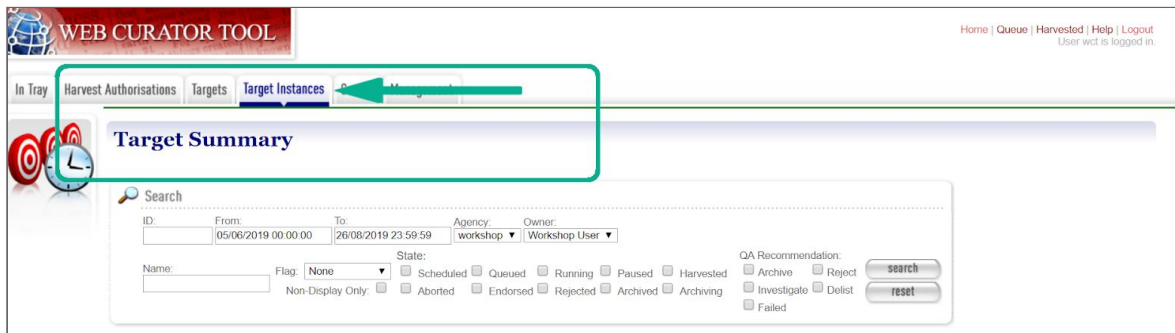
10. Save the Target. The future crawls for this Target will now be scheduled.



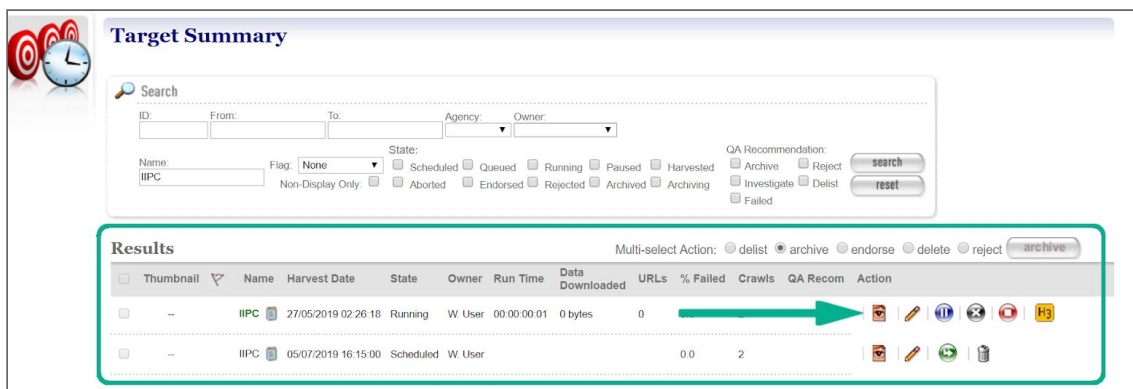
Monitor the Target Instance

The harvest that is scheduled from the Target is called a Target Instance. These can run on a pre-defined schedule or be started manually. Once they are running, there are runtime statistics and logs to assist in monitoring their progress.

1. Navigate to the Target Instance tab



2. The Target we saved earlier should have scheduled a new Target Instance to start five minutes into the future. Observe the scheduled Target Instance move into a Running state. Click the View icon.



3. Navigate to the Harvest State tab, and observe the crawl statistics.

You can see:

- how many URLs have been downloaded so far, and how many are still in the queue.
- the current and average data and URLs downloaded per second.
- the total data downloaded and crawl time elapsed.

Target Summary
IIPC (524288)

General Profile **Harvest State** [View](#) [Download](#) [Actions](#) [Display](#)

WCT Application Version: 2.0.2
Capture System: Heritrix 3.4.0
Harvest Server: H3 Agent
Job: 524288
Status: RUNNING
Average KB/s: 0
Current KB/s: 0
Average URIs/s: 0.04
Current URIs/s: 0.11
URLs Downloaded: 2
URLs Queued: 2
URLs Failed: 0
Data Downloaded: 420 bytes
Elapsed Time: 00:00:00.47

[done](#) [edit](#)

4. Navigate to the Logs tab, and click on the View link for crawl.log.

Target Summary
IIPC (524288)

General Profile Harvest State **Logs** [View](#) [Download](#) [Actions](#) [Display](#)

Filename	Action	Size
crawl.log	View	
nontatal-errors.log	View Download	0 bytes
alerts.log	View Download	0 bytes
uri-errors.log	View Download	0 bytes
progress-statistics.log	View Download	1.18 KB
runtime-errors.log	View Download	0 bytes

[done](#) [edit](#)

5. Observe the output from the Heritrix crawl log.

You can see all the URLs that Heritrix has attempted to download, successfully or otherwise.

Log viewer: crawl.log

```
2019-05-27T02:26:23.427z 1 55 dns:netpreserve.org F http://netpreserve.org/ text/dns #001 20190527022617467+5425 sha1:RMVZHNK6h5B7JVF5F6D552L7G89T-
2019-05-27T02:26:56.759z 200 67 http://netpreserve.org/robots.txt F http://netpreserve.org/ text/plain #021 20190527022651361+5370 sha1:5S2086Lk7EH0741
2019-05-27T02:27:29.393z 1 74 dns:www.googletagmanager.com EP https://www.googletagmanager.com/gtag/js?id= text/dns #009 20190527022729373+14 sha1:4H
2019-05-27T02:27:29.405z 1 46 dns:s.w.org EP http://s.w.org/ text/dns #002 20190527022729383+10 sha1:3WVYZDRGULFCNKJAD4H06IARJL4LFOY - -
2019-05-27T02:27:29.404z 1 66 dns:fonts.googleapis.com EP http://fonts.googleapis.com/ text/dns #003 20190527022729381+9 sha1:RAZ5W42AY5H52NF7AY483W
2019-05-27T02:27:29.487z 1 50 dns:theme.co EP http://theme.co/media/bg-integrity-91.jpg text/dns #017 20190527022729391+95 sha1:NS4UHRK350KGDY2D35I
2019-05-27T02:27:30.058z 200 103263 http://netpreserve.org/ - text/html #021 20190527022730058+4068 sha1:GDG304XSVOUTQXESD7E0XWNGSLEIWEV - 3t
2019-05-27T02:27:32.581z 200 237 http://s.w.org/robots.txt EP http://s.w.org/ text/plain #009 20190527022732438+102 sha1:Y4TN87Qc5JDI1YIXITZDW3ENWHQEFW
2019-05-27T02:27:32.647z 200 25 http://fonts.googleapis.com/robots.txt EP http://fonts.googleapis.com/ text/plain #002 20190527022732436+205 sha1:PIBO
2019-05-27T02:27:33.328z 404 1571 https://www.googletagmanager.com/robots.txt EP https://www.googletagmanager.com/gtag/js?id= text/html #003 201905270227
2019-05-27T02:27:35.693z 301 0 http://s.w.org/ E http://netpreserve.org/ text/html #009 20190527022735599+92 sha1:3142R35RMNFZMSVX7KXIA3CX5QmJ3 -
2019-05-27T02:27:35.953z 301 194 http://theme.co/robots.txt EP http://theme.co/media/bg-integrity-91.jpg text/html #017 2019052702273516+3426 sha1:WHP
2019-05-27T02:27:36.169z 404 1603 http://fonts.googleapis.com/ E http://netpreserve.org/ text/html #002 20190527022735653+500 sha1:YUHW5YSYRHZK4J3UXC4
2019-05-27T02:27:38.857z 200 65720 https://www.googletagmanager.com/gtag/js?id= E http://netpreserve.org/ application/javascript #017 20190527022737995+4
2019-05-27T02:27:38.982z 200 4286 http://s.w.org/favicon.ico EI http://s.w.org/ image/x-icon #017 20190527022738859+100 sha1:H4UGQXMPVHQ3DKVLEXTNORH8NK
2019-05-27T02:27:39.003z 1 55 dns:www.gstatic.com EP http://www.gstatic.com/wcm/loader.js text/dns #017 20190527022739099+13 sha1:LB00EY7B0I7VUHF
2019-05-27T02:27:39.027z 1 66 dns:adservice.google.com EXP https://adservice.google.com/dm/regclk text/dns #017 20190527022739017+8 sha1:RM3K35JE36
2019-05-27T02:27:39.045z 1 63 dns:ad.doubleclick.net EXP http://ad.doubleclick.net/activity text/dns #017 20190527022739033+10 sha1:RDLXFSERNB3PIFQCI
2019-05-27T02:27:39.069z 1 66 dns:www.googletraveladservices.com EXP https://www.googletraveladservices.com/travel/flights/clk text/dns #017 20190527
2019-05-27T02:27:39.088z 1 64 dns:www.googleadservices.com EXP http://www.googleadservices.com/pagead/conversion_async.js text/dns #017 201905270227
2019-05-27T02:27:39.116z 1 74 dns:www.google-analytics.com EXP https://www.google-analytics.com/analytics.js text/dns #021 20190527022739103+11 sha1
2019-05-27T02:27:39.400z 200 1391 http://fonts.googleapis.com/css?family=montserrat&3A400&2C700&47C00&3A100&2C300&2C400&2C700&ver=5.2.1.E http://netp
2019-05-27T02:27:39.424z 1 66 dns:fonts.gstatic.com EP http://fonts.gstatic.com/s/montserrat/v13/JTUSjIgl_16t8KChKm459Wlhgz.ttf text/dns #002 20190
2019-05-27T02:27:42.373z 404 1572 https://www.googletagmanager.com/favicon.ico EI https://www.googletagmanager.com/gtag/js?id= text/html #017 2019052702
2019-05-27T02:27:42.406z 200 2238 http://www.gstatic.com/robots.txt EXP http://www.gstatic.com/wcm/loader.js text/plain #017 20190527022742169+207 sha1:1
```

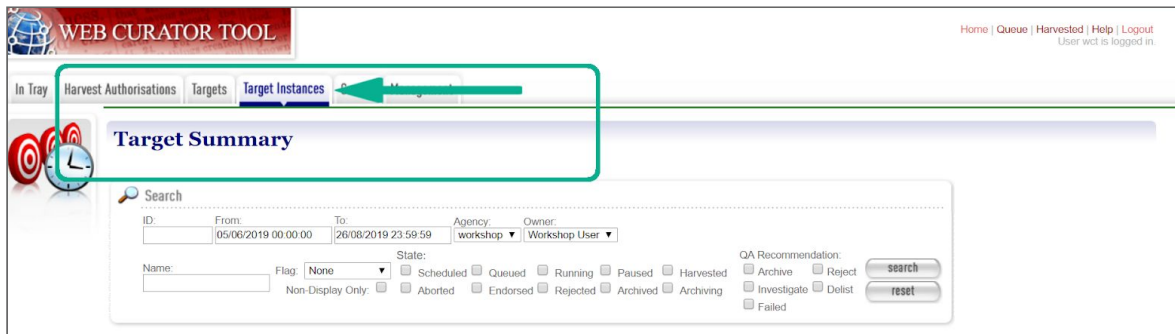
Show Line Numbers
Number of lines to display
Filter Type
[apply](#)

Quality review

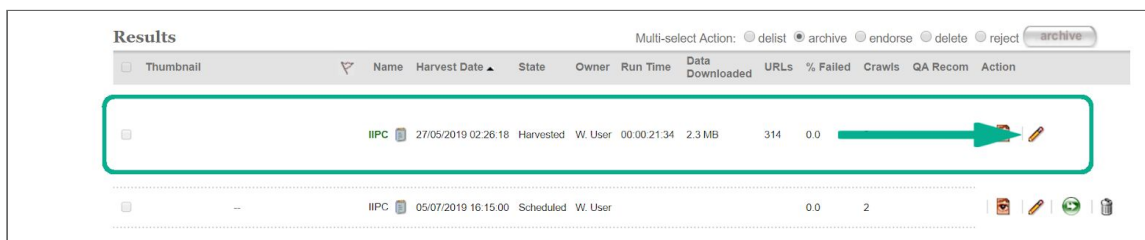
Once a Target Instance has been harvested, the next step is to review the quality of the content captured. This is primarily to determine whether the harvest is of sufficient quality to preserve and archive.

For more information on WCT quality review, see the [documentation](#).

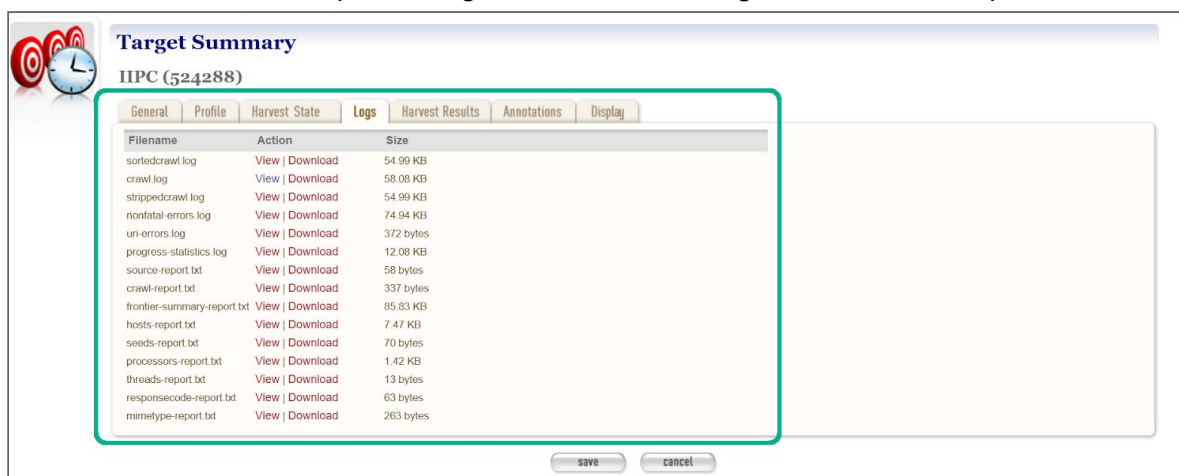
1. Navigate to the Target Instance tab



2. When the running Target Instance has finished and moved into the Harvested state, click the Edit icon.



3. Navigate to the Logs tab, and observe the now larger list of available Heritrix logs. These additional reports are generated after a Target Instance is completed.



- The hosts-report will show a breakdown of the number of URLs and bytes crawled per hostname. Once you are finished viewing a log or report, click the Done button.

Log viewer: hosts-report.txt

```
[#urls] [#bytes] [host] [#robots] [#remaining] [#novel-urls] [#novel-bytes] [#dup-by-hash-urls] [#dup-by-hash-bytes] [#not-modified-urls] [#not-modified-bytes]
114 6449 dns: 0 22 114 6449 0 0 0
12 16571 fonts.googleapis.com 0 2 12 16571 0 0 0
10 208617 www.googletagmanager.com 0 0 10 208617 0 0 0
6 122289 netpreserve.org 0 246 6 122289 0 0 0
6 4433 theme.co 0 6 4433 0 0 0
5 6159 s.w.org 0 16 5 6159 0 0 0
3 42794 wordpress.org 0 1 3 42794 0 0 0
3 5518 www.googletravelservices.com 0 0 3 5518 0 0 0
2 2296 adservice.google.com 0 0 2 2296 0 0 0
2 34563 sf.wordpress.org 0 0 2 34563 0 0 0
2 33652 ak.wordpress.org 0 0 2 33652 0 0 0
2 33817 am.wordpress.org 0 0 2 33817 0 0 0
2 37432 ar.wordpress.org 0 0 2 37432 0 0 0
2 33957 arg.wordpress.org 0 0 2 33957 0 0 0
2 33899 as.wordpress.org 0 0 2 33899 0 0 0
2 33717 ast.wordpress.org 0 0 2 33717 0 0 0
2 33833 bcc.wordpress.org 0 0 2 33833 0 0 0
2 37446 bg.wordpress.org 0 0 2 37446 0 0 0
2 40264 bo.wordpress.org 0 0 2 40264 0 0 0
2 33773 brs.wordpress.org 0 0 2 33773 0 0 0
2 34143 bs.wordpress.org 0 0 2 34143 0 0 0
2 34980 ca.wordpress.org 0 0 2 34980 0 0 0
2 33591 ceb.wordpress.org 0 0 2 33591 0 0 0
2 34840 cl.wordpress.org 0 0 2 34840 0 0 0
2 34216 co.wordpress.org 0 0 2 34216 0 0 0
```

Show Line Numbers

Number of lines to display

Filter Type

- The mimetype-report will show a breakdown of the number of URLs and bytes crawled per mime type.

Log viewer: mimetype-report.txt

```
[#urls] [#bytes] [mime-types]
114 6449 text/dns
102 36494 text/plain
78 2034365 text/html
7 10054 text/css
3 198720 application/javascript
3 10491 image/x-icon
2 2590 application/css+xml
2 16783 image/jpeg
1 46780 font/ttf
1 290 image/gif
1 44915 text/javascript
Displaying: 100% of 263 B
```

- Navigate to the Harvest Results tab, and click the Review link.
If the option is not available yet, WCT may still be indexing the harvested result. Check back in a few minutes.

In Tray Harvest Authorisations Targets **Target Instances** Groups Management

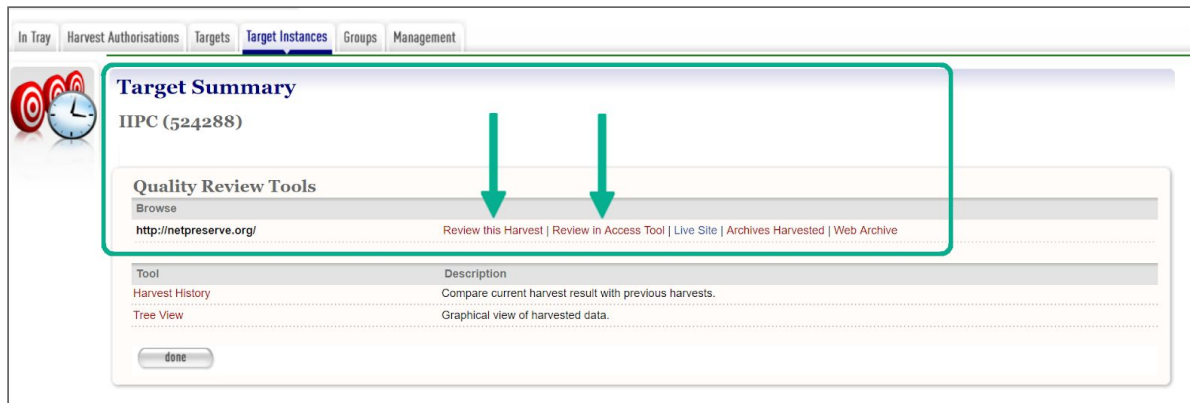
Target Summary

IIPC (524288)

General Profile Harvest State Logs **Harvest Results**

No.	Date	Derived From	User	Notes	State	Action
1	27/05/2019 02:48:04		W User	Original Harvest		Review

- There are two options for viewing the harvested result. The 'Review this Harvest' link will open the harvest in an old built-in web harvest viewer. The 'Review in Access Tool' link will open the harvest in a configured instance of OpenWayback. Click the Review in Access Tool link.

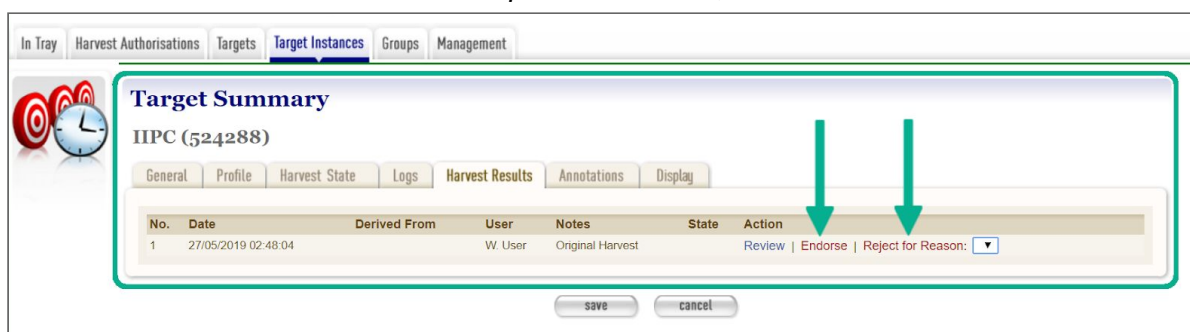


- You can browse and review the harvest here.



- Navigate back to the Harvest Results tab. Here you can choose to 'Endorse' the harvest if you are satisfied with the result. Or reject the harvest if you are not satisfied with the result. Reject reasons can be setup in the Management tab.

For more information on review process in WCT, see the [documentation](#).



Tutorial B - Crawl Scoping

What you will learn [description, profiles, scheduling, crawl monitoring, quality review]

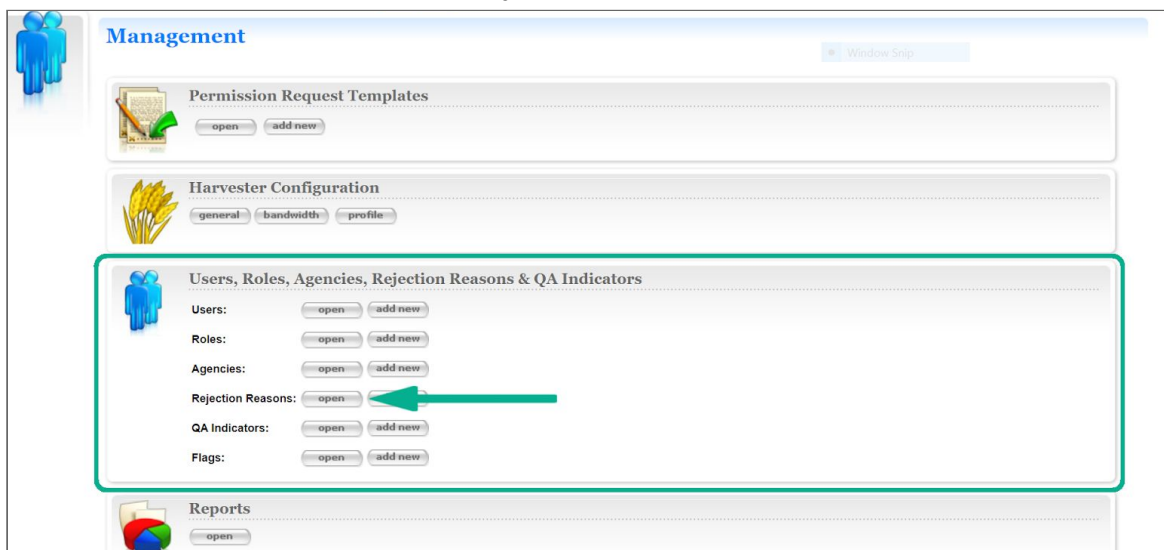
Scenario: You must crawl a website that is under a domain that your institution has authorisation to harvest. Eg. ccTLD like .uk, .eu, .nz or institution like harvard.edu

Setup Rejection Reasons

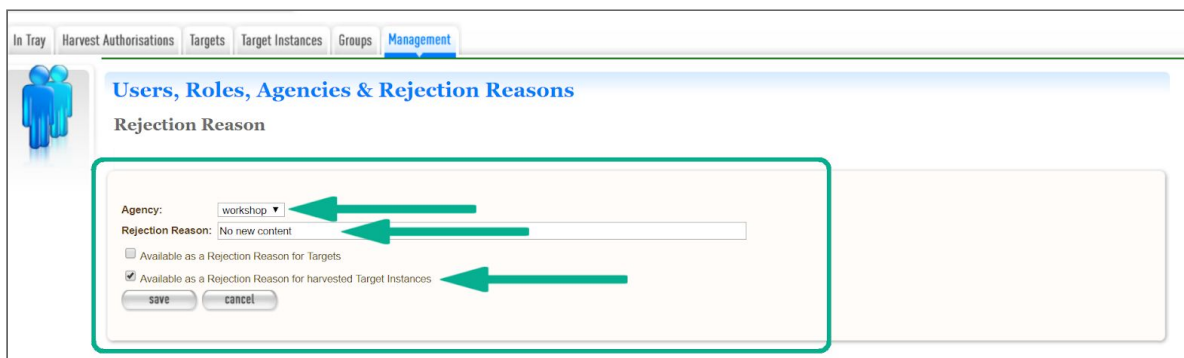
A rejection reasons can be assigned to Targets and Target Instances to state why it has been rejected. For example, a Target might be rejected for curatorial reasons, or a harvested Target Instance might be rejected for technical reasons.

For more information on Rejection Reasons in WCT, see the [documentation](#).

1. Navigate to the Management tab in WCT.
2. Click the Open option next to Rejection Reasons.



3. Create a new Rejection Reason, and enter the following detail, and click Save.
Agency: workshop
Rejection Reason: No new content
Available as a Rejection Reason for harvested Target Instances: <tick>



4. Create another Rejection Reason, and enter the following detail, and click Save.
Agency: workshop
Rejection Reason: Technical reasons
Available as a Rejection Reason for harvested Target Instances: <tick>

The rejection reason 'Technical reasons' has been created successfully.

Rejection Reason	Available for Targets	Available for Target Instances	Agency	Action
No new content	No	Yes	workshop	
Technical reasons	No	Yes	workshop	

Setup Target

A Target is a record that describes the thing you intend to harvest from the Web. It contains metadata descriptions, seed URLs, how and when the thing will be crawled.

For more information on WCT Targets, see the [documentation](#).

1. Navigate to the Targets tab in WCT.
2. Create a new Target.
3. In the General tab enter the following detail
Name: <short name of your institution>
Description: <full name of your institution> website
State: Approved

Targets

General Seeds Profile Schedule Annotations Description Groups Access

Id:

Name: NLNZ *

Description: National Library of New Zealand website

Reference Number:

Run on Approval:

Use Automated QA:

Owner: Workshop User

State: Approved

Auto-prune:

Reference Crawl:

Request to Archivists:

- Navigate to the Seeds tab. If the Authorisation option is set to auto, then any new seeds will automatically be linked to Harvest Authorisations with matching URL patterns. Enter the following detail, then click the Link button.

Seed: <your institutional website>

Authorisation: Auto

Targets
NLNZ

General Seeds Profile Schedule Annotations Description Groups Access

Seed:

Authorisation: Auto

Seed	Primary	Harvest Auth	Auth Agent	Start	End	Status	Action
<input type="checkbox"/> https://natlib.govt.nz/	<input checked="" type="checkbox"/>	Blanket Authorisation	Dummy Authorising Agency	01/05/2019	30/04/2029	Approved	<input type="button" value="add"/> <input type="button" value="edit"/> <input type="button" value="delete"/>

- Navigate to the Profile tab, and ensure the following detail is set.

Harvester Type: HERITRIX3

Base Profile: Default - workshop

- Navigate to the Description tab, and enter some descriptive metadata for your institution website. The following data is an example.

Description: <full name of your institution>

Subject: GLAM

Type: Collection

Language: <language of your institution website>

Targets
NLNZ

General Seeds Profile Schedule Annotations Description Groups Access

Title:

Identifier:

Description:

Subject:

Creator:

Publisher:

Contributor:

Type:

Format:

Source:

Language:

Relation:

Coverage:


ISSN:

ISBN:

- Navigate to the Schedule tab, and tick the Harvest Now box.
This will schedule a new Target Instance to start in five minutes.

Targets
NLNZ

General Seeds Profile **Schedule** Annotations Description Groups Access

Harvest Now: 
 Allow harvest optimization:

create new

Schedule	Owner	Next Scheduled Time	Action
save cancel			

- Save the Target.

Monitor Target Instance

The harvest that is scheduled from the Target is called a Target Instance. These can run on a pre-defined schedule or be started manually. Once they are running, there are runtime statistics and logs to assist in monitoring their progress.

- For the newly created Target, click the View Target Instances icon. This will search for all the Target Instances for this Target.


Targets

Target 'NLNZ' has been saved. A new instance has been scheduled to run in 5 minutes!

Search

ID: Name: Seed: Agency: User: Sort Order:
 Description: Member of: Non-Display Only State: Pending Reinstated Nominated Rejected
 Approved Cancelled Completed search reset


Results create new

Id	Created	Name	Agency	Owner	Status	Seeds	Action
425984	27/05/19	IIPC	workshop	Workshop User	Approved	http://netpreserve.org/	
622592	28/05/19	NLNZ	workshop	Workshop User	Approved	https://natlib.govt.nz/	

Results 1 to 2 of 2
 Page 1 of 1
 Rows per page: 10

- Observe the scheduled Target Instance, and click the Harvest Now icon.

Results Multi-select Action: delist archive endorse delete reject archive

Thumbnail	Name	Harvest Date	State	Owner	Run Time	Data Downloaded	URLs	% Failed	Crawls	QA Recom	Action
	NLNZ	28/05/2019 00:48:53	Scheduled	W. User				0.0	1		

Results 1 to 1 of 1
 Page 1 of 1
 Rows per page: 10

3. Assign the Target Instance to the available Harvest Agent, by clicking Allocate.

Target Summary
NLNZ (688129)

Harvest Now
Id: 688129
Target Name: NLNZ
Schedule: 28/05/2019 00:57:50

Harvest Agent	Max Harvests	Current Harvests	
H3 Agent	2	0	allocate

done

4. Observe the scheduled Target Instance move into a Running state. Click the View icon.

Target Summary

Search
ID: From: 28/05/2019 10:00:00 To: Agency: Owner:

Name: NLNZ Flag: None State: Scheduled Queued Running Paused Harvested QA Recommendation: Archive Reject search
Non-Display Only: Aborted Endorsed Rejected Archived Archiving Investigate Delist reset
Failed

Results Multi-select Action: delist archive endorse delete reject archive

Thumbnail	Name	Harvest Date	State	Owner	Run Time	Data Downloaded	URLs	% Failed	Crawls	QA Recom	Action
	NLNZ	28/05/2019 10:05:50	Running	W User	00:00:00:00	0 bytes	0				View

Results 1 to 1 of 1
Page 1 of 1
Rows per page: 10

5. Navigate to the Harvest State tab, and observe the crawl statistics. These statistics give a high-level overview of the running crawl.

Certain behaviours can be a red flag for a bad crawl. Watch out for:

- a high proportion of failed URLs,
- the amount of URLs or Data downloaded is not increasing

Target Summary
NLNZ (753665)

General Profile **Harvest State** Logs Harvest Results Annotations Display

WCT Application Version: 2.0.2
Capture System: Heritrix 3.4.0
Harvest Server: H3 Agent
Job: 753665
Status: RUNNING
Average KB/s: 17
Average URI/s: 0.72
Current KB/s: 32
Current URI/s: 3.31

URLs Downloaded: 89
URLs Queued: 343
URLs Failed: 0
Data Downloaded: 2.06 MB
Elapsed Time: 00:00:02:04

- Navigate to the Logs tab, and click on the View link for crawl.log.
- Scroll through the crawl log output. For this tutorial, locate any URLs from a hostname that you did not expect or want to be included in this crawl.

For example, URLs from the Pinterest host.

Log viewer: crawl.log

```

2019-05-28T10:07:55.445Z 301 194 http://embed.ly/code?url=LREX https://cdn.embedly.com/widgets/platform.js text/html #019
2019-05-28T10:07:55.446Z 200 5093 https://twitter.com/intent/tweet LREX https://cdn.embedly.com/widgets/platform.js text/html
2019-05-28T10:07:55.464Z 302 0 https://www.pinterest.com/pin/create/button/ LREX https://cdn.embedly.com/widgets/platform
2019-05-28T10:07:56.010Z 301 76 dns:www.pinterest.nz LREXRP https://www.pinterest.nz/pin/create/button/ text/dns #014 2019
2019-05-28T10:07:56.487Z 200 185 https://api-cdn.embedly.com/ LREX https://cdn.embedly.com/widgets/platform.js text/html #007
2019-05-28T10:07:58.362Z 200 4632 https://natlib.govt.nz/kakapo/assets/favicon-cal352313f0ac0a33b4934eb70a5f05c4376f93c6a10a
2019-05-28T10:07:58.453Z 200 25726 https://images.ctfassets.net/n3p1o4aq7k6m/2z51eE2zm4ACMcCe6oiOI/3ac0c4465fad4d0e4e95e00a8
2019-05-28T10:07:58.463Z 200 706 https://www.youtube.com/robots.txt LREXRP https://www.youtube.com/watch?v= text/plain #015
2019-05-28T10:07:58.509Z 200 2656 https://app.embed.ly/ LREX https://cdn.embedly.com/widgets/platform.js text/html #005 2019
2019-05-28T10:07:58.566Z 200 1037 https://vimeo.com/robots.txt LREXRP https://vimeo.com/ text/plain #020 20190528100758112+3
2019-05-28T10:07:59.077Z 200 184 http://cdn.embedly.com/widgets/embed?url=LREX https://cdn.embedly.com/widgets/platform.js
2019-05-28T10:07:59.466Z 404 9377 https://www.pinterest.nz/robots.txt LREXRP https://www.pinterest.nz/pin/create/button/ text
2019-05-28T10:07:59.702Z 200 178 https://embed.ly/robots.txt LREXRP https://embed.ly/code?url= text/html #019 2019052810075
2019-05-28T10:08:01.455Z 200 4632 https://natlib.govt.nz/kakapo/assets/favicon-cal352313f0ac0a33b4934eb70a5f05c4376f93c6a10a
2019-05-28T10:08:01.792Z 404 21256 https://soundcloud.com/track/ LREX https://cdn.embedly.com/widgets/platform.js text/html #
2019-05-28T10:08:01.849Z 301 60 https://images.ctfassets.net/favicon.ico LREX https://images.ctfassets.net/n3p1o4aq7k6m/1F
2019-05-28T10:08:02.068Z 200 0 https://www.youtube.com/watch?v=LREX http://www.youtube.com/watch?v= text/html #015 2019
2019-05-28T10:08:02.130Z 200 25546 https://vimeo.com/ LREXR http://vimeo.com/ text/html #020 20190528100801514+527 sha1:27282
2019-05-28T10:08:02.190Z 403 243 https://cdn.embedly.com/widgets/ LREX https://cdn.embedly.com/widgets/platform.js applicat
2019-05-28T10:08:02.951Z 301 0 https://www.pinterest.nz/pin/create/button/ LREXR https://www.pinterest.com/pin/create/bu
2019-05-28T10:08:05.539Z 200 241 http://natlib.govt.nz/robots.txt LXP http://natlib.govt.nz/kakapo/assets/wellsprint_logo-c
2019-05-28T10:08:06.062Z 200 13825 https://embed.ly/code?url=LREXR http://embed.ly/code?url= text/html #025 2019052810080452
2019-05-28T10:08:07.767Z 200 - https://www.pinterest.nz/login/?next=/pin/create/button/ LREXRP https://www.pinterest.nz/p
4632 http://natlib.govt.nz/kakapo/assets/wellsprint_logo-cal352313f0ac0a33b4934eb70a5f05c4376f9
546371 https://www.youtube.com/ LREXRR https://www.youtube.com/watch?v= text/html #015 20190528100

```

Show Line Numbers

Number of lines to display

Filter Type

- To filter the crawl log output, and only see entries for your unwanted hostname, set the Filter Type to 'Return all lines matching regex', enter the hostname surrounded by the any character pattern `.*` and click Apply.

For example, `.*pinterest.*`

Log viewer: crawl.log

```

2019-05-28T10:07:48.106Z 1 76 dns:www.pinterest.com LREXP https://www.pinterest.com/pin/create/button/ text/dns #016 201905
2019-05-28T10:07:48.200Z 1 76 dns:log.pinterest.com LREXP https://log.pinterest.com/ text/dns #016 20190528100748159+41 sha
2019-05-28T10:07:51.904Z 200 0 https://log.pinterest.com/robots.txt LREXP https://log.pinterest.com/ unknown #003 2019052810
2019-05-28T10:07:52.026Z 302 0 https://www.pinterest.com/robots.txt LREXP https://www.pinterest.com/pin/create/button/ text/
2019-05-28T10:07:55.305Z 200 0 https://www.pinterest.com/ LREX https://cdn.embedly.com/widgets/platform.js unknown #003 2019
2019-05-28T10:07:55.464Z 302 0 https://www.pinterest.com/pin/create/button/ LREX https://cdn.embedly.com/widgets/platform.js
2019-05-28T10:07:55.537Z 1 76 dns:www.pinterest.nz LREXRP https://www.pinterest.nz/pin/create/button/ text/dns #014 201905
2019-05-28T10:07:59.072Z 200 9377 https://www.pinterest.nz/robots.txt LREXRP https://www.pinterest.nz/pin/create/button/ text/p
2019-05-28T10:08:02.580Z 302 0 https://www.pinterest.nz/pin/create/button/ LREXR https://www.pinterest.com/pin/create/butto
2019-05-28T10:08:05.583Z -9998 - https://www.pinterest.nz/login/?next=/pin/create/button/ LREXRR https://www.pinterest.nz/pin/

```

Displaying: 2.2% of 89 KB

Show Line Numbers

Number of lines to display

Filter Type

Regular Expression

- You should now only see lines in the crawl log relevant to the hostname you identified.

For more information on the WCT log file viewer, see the [documentation](#).

- Navigate back to the Target Instance tab, tick the Running state filter, and click Search.

The screenshot shows the 'Target Summary' page with a search bar. The search filters include ID, From, To, Agency, and Owner. The 'Running' state filter is selected. The search button is highlighted with a green box and a green arrow. Below the search bar, there are checkboxes for various states: Scheduled, Queued, Running, Paused, Harvested, Aborted, Endorsed, Rejected, Archived, and Archiving. The 'search' button is located to the right of these filters.

- Ensure you can see your running Target Instance, and click the Stop icon. *After clicking Stop, the state can sometimes take up to a minute to change from Running to Stopping. WCT is waiting on Heritrix to shutdown the crawl job and send back a new status update.*

The screenshot shows the 'Target Summary' page with a table of results. The table has columns for Thumbnail, Name, Harvest Date, State, Owner, Run Time, Data Downloaded, URLs, % Failed, Crawls, QA Recom, and Action. The 'Running' state is selected. The 'Stop' icon in the Action column is highlighted with a green box and a green arrow.

Thumbnail	Name	Harvest Date	State	Owner	Run Time	Data Downloaded	URLs	% Failed	Crawls	QA Recom	Action
	NLNZ	28/05/2019 10:05:50	Running	W. User	00:00:30:34	46.04 MB	889	0.0	3		Stop, Edit, Refresh, Delete, H3

- Stopping the Target Instance will keep the harvested content for review. Aborting the Target Instance will discard the harvested content and will not be available for review.

Adjust the Target

We want to adjust the Target Profile settings in order to exclude URLs from our unwanted hostname. We also want to try and restrict our crawl from searching too far away from the hostname for our seed URL.

- Navigate to the Targets tab.
- Click the Edit icon of the Target for your institution website.

The screenshot shows the 'Results' page with a table of targets. The table has columns for Id, Created, Name, Agency, Owner, Status, Seeds, and Action. The target with ID 622592 is highlighted with a green box, and the 'Edit' icon in the Action column is highlighted with a green arrow.

Id	Created	Name	Agency	Owner	Status	Seeds	Action
425984	27/05/19	IIPC	workshop	Workshop User	Approved	http://netpreserve.org/	Stop, Edit, Refresh, Delete, H3
622592	28/05/19	NLNZ	workshop	Workshop User	Completed	https://natlib.govt.nz/	Stop, Edit, Refresh, Delete, H3

- Navigate to the Profile tab. We will exclude URLs from our unwanted hostname in the new crawl and reduce the number of 'hops' crawled from the Target seed URL. Enter the following detail, and tick the corresponding *Enable Override* checkbox.

Max Hops: 4

*Exclusion filter: .*pinterest.**

Targets
NLNZ

General Seeds **Profile** Schedule Annotations Description Groups Access

Base Profile
Harvester Type: HERITRIX3
Base Profile: Default - workshop

Profile Element	Override Value	Enable Override
Document Limit		<input type="checkbox"/>
Data Limit	0.0 B	<input type="checkbox"/>
Time Limit	0.0 SECOND	<input type="checkbox"/>
Max Path Depth		<input type="checkbox"/>
Max Hops	4	<input checked="" type="checkbox"/>
Max Transitive Hops		<input type="checkbox"/>
Ignore Robots.txt	<input type="checkbox"/>	<input type="checkbox"/>
Ignore Cookies	<input type="checkbox"/>	<input type="checkbox"/>
Extract Javascript	<input type="checkbox"/>	<input type="checkbox"/>
Block URLs	.*pinterest.*	<input checked="" type="checkbox"/>
Include URLs		<input type="checkbox"/>

- Navigate to the Schedule tab, tick the Harvest Now box, and click Save.
- Click the View Target Instances icon next to the Target.
- Start the Target Instance using the Harvest Now option.

Compare and review the crawl

- View the running Target Instance. Monitor the crawl log to ensure you don't see any URLs from your now excluded hostname. Filter on the regular expression
*Regular Expression: .*pinterest.**

Log viewer: crawl.log

Displaying: 0% of 67 KiB

Show Line Numbers

Number of lines to display 700

Filter Type Return all lines matching regex

Regular Expression .*pinterest.*

apply

- Let the Target Instance run for 5-10 minutes, or long enough to be sure that the exclusion filter has worked, before manually stopping it.

- Once the Target Instance has moved into the *Harvested* state, click on the Target Instance name to go the the Target Summary screen.

Search

ID: 1048576 From: To: Agency: Owner:

Name: NLNZ Flag: None State: [checkbox] Scheduled [checkbox] Queued [checkbox] Running [checkbox] Paused [checkbox] Harvested [checkbox] Archive [checkbox] Reject [checkbox] Investigate [checkbox] Delist [checkbox] Failed

Multi-select Action: [radio] delist [radio] archive [radio] endorse [radio] delete [radio] reject **archive**

Thumbnail	Name	Harvest Date	State	Owner	Run Time	Data Downloaded	URLs	% Failed	Crawls	QA Recom	Action
	NLNZ	31/05/2019 09:39:19	Harvested	W. User	00:00:09:49	12.32 MB	273	0.0	4		

Results 1 to 1 of 1
Page 1 of 1
Rows per page: 10

- The Target Summary has several areas to assist with quality review. The Harvest History table shows comparable statistics for the previous crawls. The Key Indicators and Recommendation sections are blank because the indicators must be setup in the Management tab.

For more information on the Target Summary, see the [documentation](#).

Target Summary

NLNZ (1048576)

Harvest Results

#	Date	Notes	State
1	31/05/2019 09:49:58	Original Harvest	Harvested

Resources

The following resources are associated with this instance:

Seeds: <https://natlib.govt.nz/>

14 Logs

Key Indicators

Indicators collected for this instance:

Recommendation

Profile Overrides

Base Profile: Default - workshop

Profile Element	Override Value	Enable Override
Ignore Robots:	<input type="checkbox"/>	<input type="checkbox"/>
Ignore Cookies:	<input type="checkbox"/>	<input type="checkbox"/>
Extract Javascript:	<input type="checkbox"/>	<input type="checkbox"/>
Max Data:	0.0 Data Unit: B	<input type="checkbox"/>
Max Docs:		<input type="checkbox"/>
Max Path Depth:		<input type="checkbox"/>
Max Hops:	4	<input checked="" type="checkbox"/>
Max Transitive Hops:		<input type="checkbox"/>
Max Time:	0.0 Time Unit: SECOND	<input type="checkbox"/>
Blocked URLs:	*pinterest.*	<input checked="" type="checkbox"/>
Included URLs:		<input type="checkbox"/>

Schedule

From	To	Type	Day/Month/Options

Annotations

Harvest History

Start Date	State	Data	URLs	Failed	Elapsed	KB/s	Job Status	QA Status
31/05/2019 09:39:19	Harvested	12.32MB	273	0	9m49s	21.0	Finished	Harvested
28/05/2019 10:05:50	Harvested	50.62MB	1022	0	36m56s	23.0	Finished	Harvested
28/05/2019 00:58:07	Aborted	75.84MB	1813	0	4h8m42s	5.0	RUNNING	Harvested
28/05/2019 00:49:14	Harvested	1.64MB	55	0	2m26s	11.0	Finished	Harvested

5. Navigate back to the Target Instance tab, and click the Edit icon for the same Target Instance.

Results

Multi-select Action: delist archive endorse delete reject

Thumbnail	Name	Harvest Date	State	Owner	Run Time	Data Downloaded	URLs	% Failed	Crawls	QA Recom	Action
<input type="checkbox"/>	NLNZ	31/05/2019 09:39:19	Harvested	W User	00:00:09:49	12.32 MB	273	0.0			

6. Navigate to the Harvest Results tab and click Review.

Target Summary

NLNZ (1048576)

General Profile Harvest State Logs **Harvest Results** Annotations Display

No.	Date	Derived From	User	Notes	State	Action
1	31/05/2019 09:49:58		W User			Review Endorse Reject for Reason: No new content

7. Click Harvest History to also see a comparison of previous crawls.

Target Summary

NLNZ (1048576)

Quality Review Tools

Browse

<https://natlib.govt.nz/> [Review this Harvest](#) | [Review in Access Tool](#) | [Live Site](#) | [Archives Harvested](#) | [Web Archive](#)

Tool	Description
Harvest History	Compare current harvest result with previous harvests.
Tree View	Graphical view of harvested data.

Tutorial C - Advanced Harvest Authorizations

What you will learn [authorizations and permissions, scheduling]

Scenario: Crawl a website that your institution is required to request permission to harvest using a WCT permission request template.

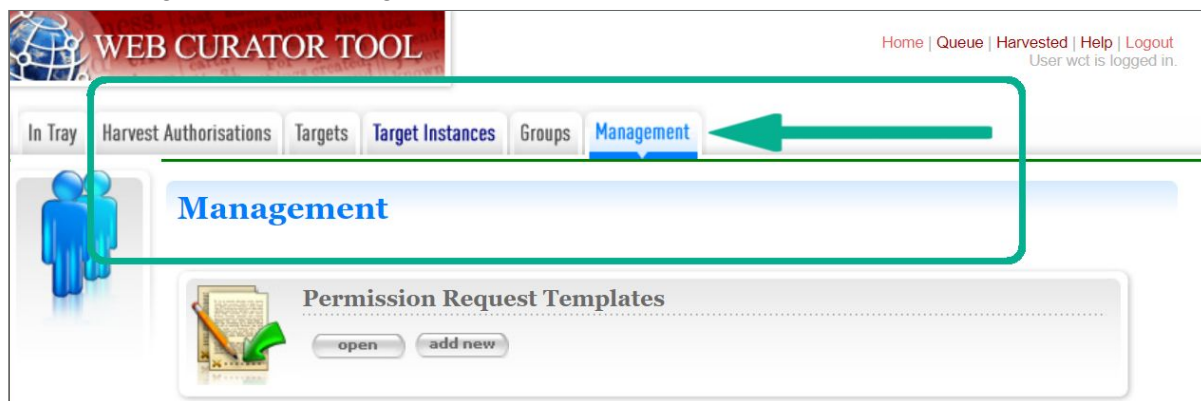
In many countries across the world deposit legislation allows libraries and other heritage institutions to harvest and store all websites belonging to their national domain. Of course what does and does not belong to the national domain will often be subject to debate, but in terms of authorisation the situation tends to be simple.

However, when a country lacks a legal deposit, web archivists are generally required to ask website owners for permission to archive their website. The Harvest Authorisations functionality of WCT was designed to deal with the workflows and record management that typically arise in that case. In this tutorial we will take a closer look at these features.

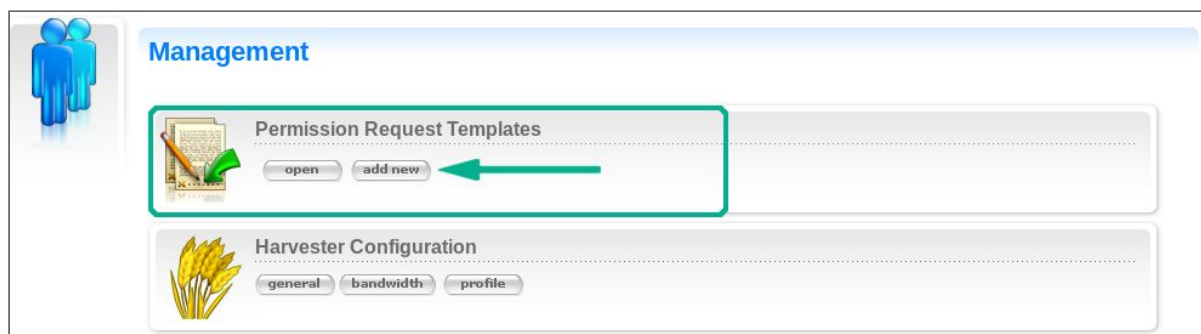
Create a Permission Request Template

First, we need to setup a permission request template that WCT will use to generate an email or letter, requesting permission to harvest a site.

1. Navigate to the Management tab.



2. Click the add new button under Permission Request Templates.



3. Select *workshop* in the Agency dropdown and enter the following values.
Template Name: Generic Permission Request Template
Template Description: This is our default template.
Template text: <the text of the email that will be sent to the owner of the site>



Permission Request Templates

Add/Edit Template

Template Name:	<input type="text" value="Generic Permission Request Template"/>	*
Template Description:	<input type="text" value="This is our default template."/>	
Agency:	<input type="text" value="workshop"/>	←
Template type:	<input type="text" value="Email Template"/>	
Template Subject:	<input type="text" value="Web Preservation Programme"/>	*
Template Overwrite From:	<input type="text" value="No"/>	
Template From Address:	<input type="text"/>	
Template CC Address(s):	<input type="text"/>	
Template BCC Address(s):	<input type="text"/>	
Template Reply-to Address:	<input type="text"/>	
Template text:	<input type="text" value="Dear Madam or Sir,

We have the pleasure to inform you that your website https://example.com is of interest to the Web Preservation Programme at the National Library of Erewhon.

We are required by law to ask for your permission to harvest and store a copy of the aforementioned website. Would you be so kind as to let us know if you consent to the harvesting and long-term storage of example.com?

Kind regards,

Jane Doe
Web Curator at the National Library of Erewhon"/>	*

Create Harvest Authorisation

Next, we're going to create a new Harvest Authorisation.

1. Navigate to the Harvest Authorisations tab.

2. Create a new Harvest Authorisation.

WEB CURATOR TOOL Home | Queue

In Tray **Harvest Authorisations** Targets Target Instances Groups Management

Harvest Authorisations

Search

ID: Name: Authorising Agent: Order No: Agency: workshop Sort Order: Name (ascending) Show Disabled:

URL Pattern: Permissions File Reference: Permissions Status: Pending Requested Approved Rejected search reset

Results

Id	Created	Name	Auth Agent	Order No	Status	Action
65536	07/05/19	Blanket Authorisation	Dummy Authorising Agency		Approved	

[create new](#)

3. Under the General tab enter:
Title: Advanced Authorisation

Harvest Authorisations

General URL Patterns Authorising Agencies Permissions

Agency: workshop

Title: *

Description:

Order No:

Published:

Enabled:

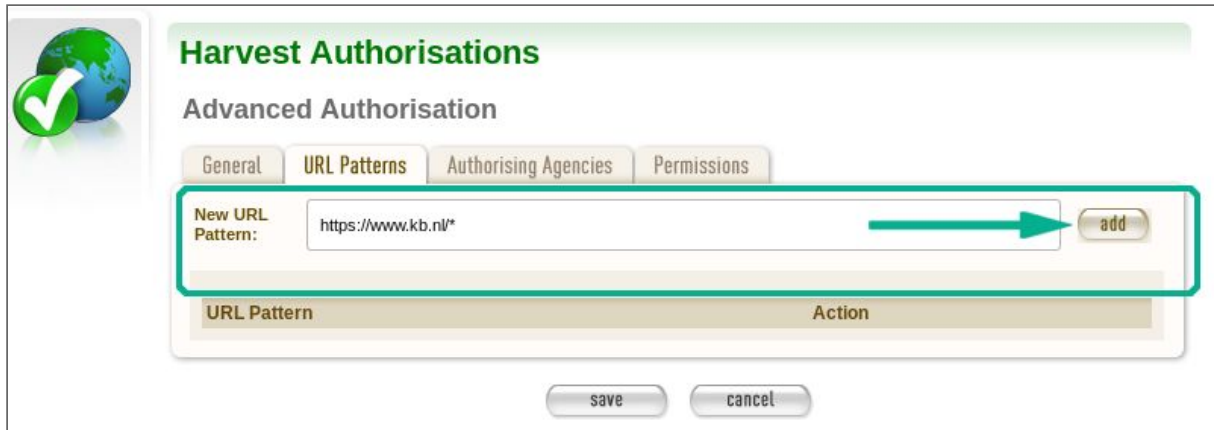
Annotations

[add](#)

Date	User	Notes	Action
There are no annotations available.			

[save](#) [cancel](#)

- In the URL pattern tab, for the purposes of this workshop, enter a pattern corresponding to your own institutional website. You can use wildcards, representing zero or more characters, by using an asterisk, e.g. `https://www.kb.nl/*`, which would cover all seeds starting with `https://www.kb.nl/`.



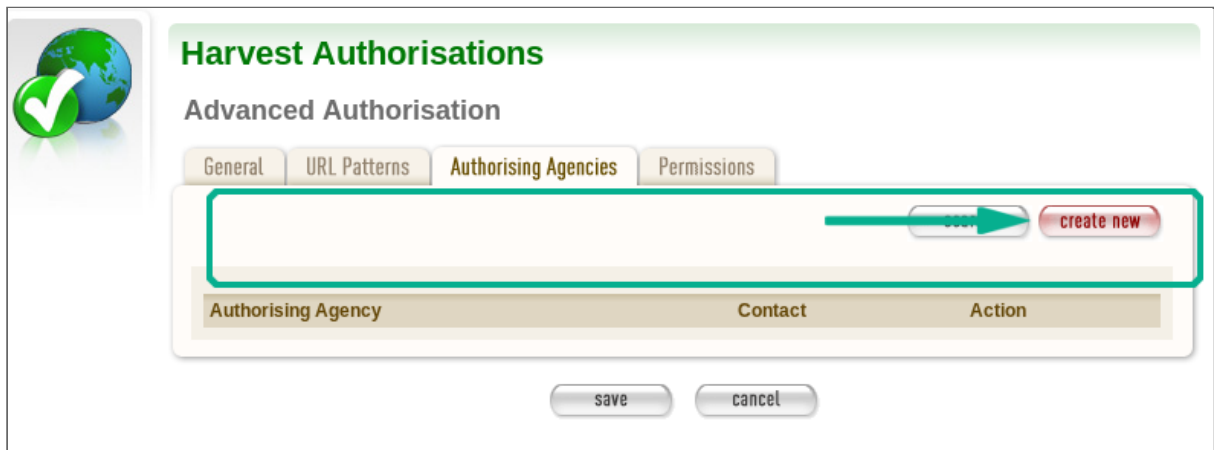
Harvest Authorisations
Advanced Authorisation

General | **URL Patterns** | Authorising Agencies | Permissions

New URL Pattern:

URL Pattern	Action
-------------	--------

- Create a new Authorising Agency under the next tab.




Harvest Authorisations
Advanced Authorisation

General | URL Patterns | **Authorising Agencies** | Permissions

Authorising Agency	Contact	Action
--------------------	---------	--------

- Fill in the contact details of the site owner. For the purposes of this workshop, use the name of your institution. The email address doesn't have to be real.



Name: *

Description:

Contact: *

Phone:

Email:

Address:

7. Create a new set of permissions under the Permissions tab.



The screenshot shows the 'Harvest Authorisations' interface. The 'Permissions' tab is selected. A red box highlights the 'create new' button, with a red arrow pointing to it. Below the tabs is a table with columns: Status, Date Requested, Authorising Agent, From, To, URL Patterns, and Action.

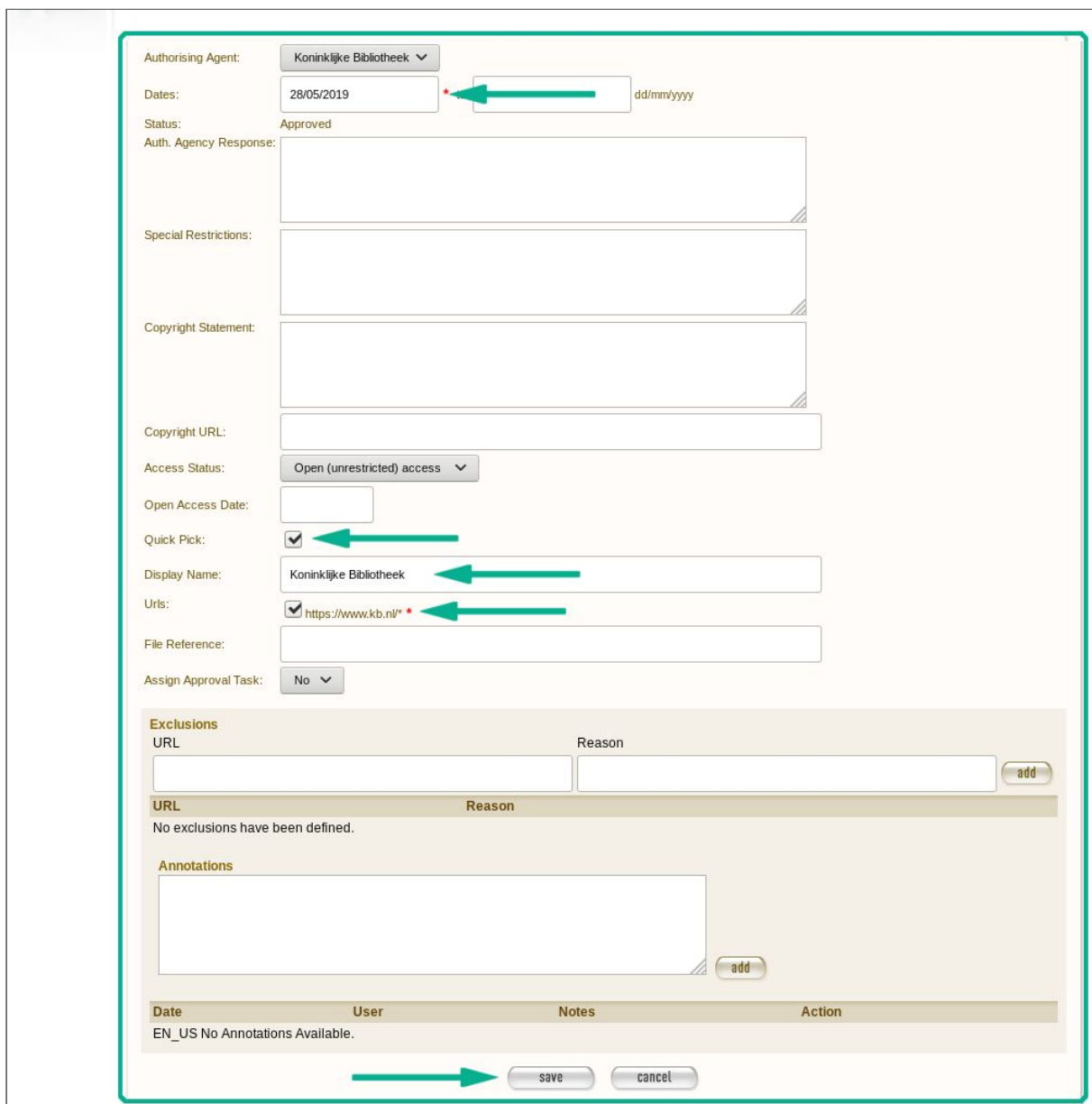
8. Enter the following detail, and click Save.

Dates: <the current date> to <blank>

Quick Pick: <tick> (this will turn out to be useful when we setup a Target)

Display Name: <name of your institution>

Urls: <tick>



The screenshot shows the 'Advanced Authorisation' form. Red arrows point to the following fields: 'Dates' (28/05/2019), 'Quick Pick' (checked), 'Display Name' (Koninklijke Bibliotheek), and 'Urls' (checked, https://www.kb.nl/*). The 'Access Status' is set to 'Open (unrestricted) access'. The 'Assign Approval Task' is set to 'No'. The 'Exclusions' section is empty, and the 'Annotations' section is also empty. The 'save' button is highlighted with a red arrow.

9. Finally, save the Harvest Authorisation.

Harvest Authorisations
Advanced Authorisation

General | URL Patterns | Authorising Agencies | **Permissions**

[create new](#)

Status	Date Requested	Authorising Agent	From	To	URL Patterns	Action
Pending		Koninklijke Bibliotheek	28/05/2019		https://www.kb.nl*	

[save](#) [cancel](#)

Send Permission Request

We are now ready to send the permission request to the website owner.

1. In the last step we returned to the list of Harvest Authorisations. Note that the status of the Harvest Authorisation is *Pending*. Now click on the letter icon behind your newly created authorisation.

Harvest Authorisations

Authorisation 'Advanced Authorisation' has been saved.

Search

ID: Name: Authorising Agent: Order No: Agency: Sort Order: Show Disabled:

URL Pattern: Permissions File Reference: Permissions Status: Pending Requested Approved Rejected

[search](#) [reset](#)

Results [create new](#)

Id	Created	Name	Auth Agent	Order No	Status	Action
229379	28/05/19	Advanced Authorisation	Koninklijke Bibliotheek		Pending	
65536	07/05/19	Blanket Authorisation	Dummy Authorising Agency		Approved	

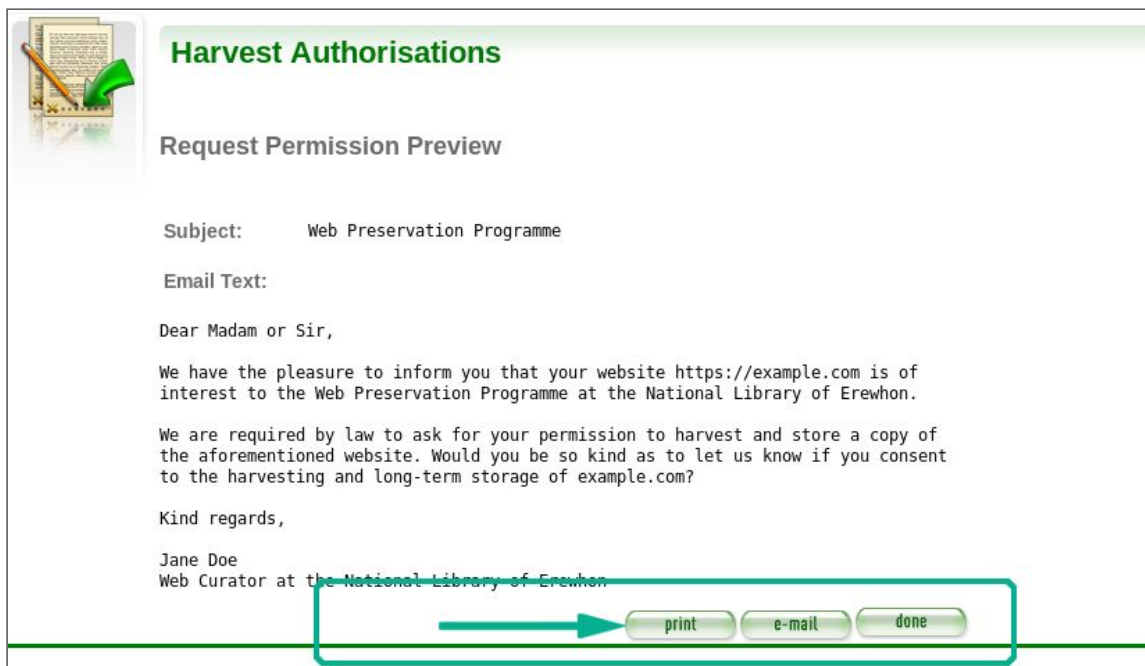
2. You are now on a screen where you can choose which request to send to the site owner. In our case there's only one permission request template. Click on the letter icon.

Harvest Authorisations

Status	Authorising Agent	From	To	URL Patterns	Action
Pending	Koninklijke Bibliotheek	28/05/2019		https://www.kb.nl*	Generic Permission Request Template

[done](#)

- This opens the text of the permission request for inspection before we send it. Since SMTP is probably not configured on the server that's running your WCT right now, we won't actually send the email, but we will **click on the print button and then cancel the ensuing print system dialog**. This will have the same effect on the status of the harvest authorisation as sending the email.



Harvest Authorisations

Request Permission Preview

Subject: Web Preservation Programme

Email Text:

Dear Madam or Sir,

We have the pleasure to inform you that your website <https://example.com> is of interest to the Web Preservation Programme at the National Library of Erewhon.

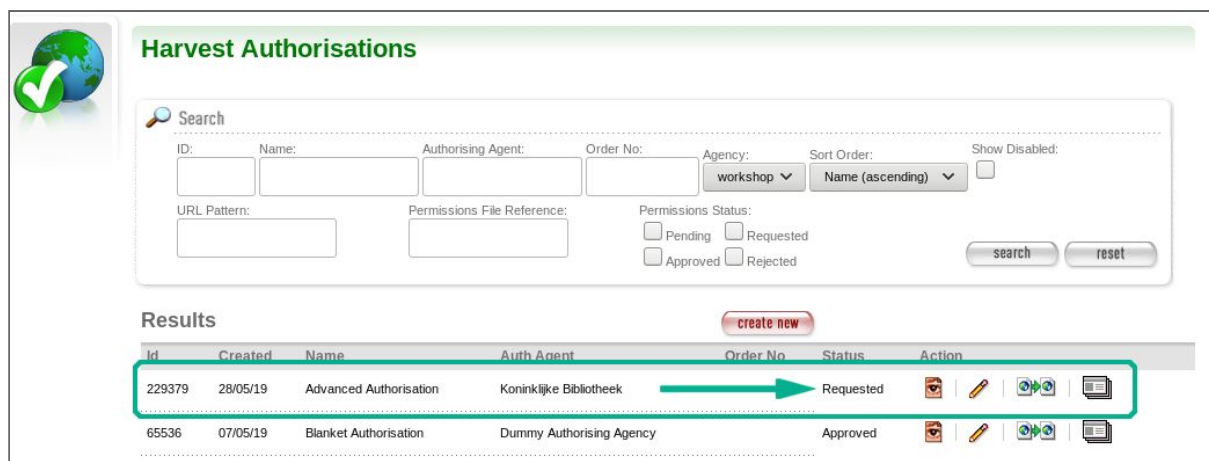
We are required by law to ask for your permission to harvest and store a copy of the aforementioned website. Would you be so kind as to let us know if you consent to the harvesting and long-term storage of example.com?

Kind regards,

Jane Doe
Web Curator at the National Library of Erewhon

[print](#) [e-mail](#) [done](#)

- Note that after printing or sending the email, the Harvest Authorisation has the status *Requested*.



Harvest Authorisations

Search

ID: Name: Authorising Agent: Order No: Agency: Sort Order: Show Disabled:

URL Pattern: Permissions File Reference: Permissions Status: Pending Requested Approved Rejected

[search](#) [reset](#)

Results [create new](#)

Id	Created	Name	Auth Agent	Order No	Status	Action
229379	28/05/19	Advanced Authorisation	Koninklijke Bibliotheek		Requested	
65536	07/05/19	Blanket Authorisation	Dummy Authorising Agency		Approved	

Revisiting the Harvest Authorisation

After some time the site owner will hopefully respond positively to your permission request. When that happens you will need to revisit the Harvest Authorisation to record the permission and change the status of the Permissions to *Approved*.

Let's assume that the owner agrees to your request, but with a few restrictions:

- they only want the archived copy to be available in the reading room of your institution;
- you're only allowed to crawl the site after some future date.

1. Navigate to the Harvest Authorisations tab.
2. Open your Advanced Authorisation in edit mode.

Results							create new
Id	Created	Name	Auth Agent	Order No	Status	Action	
589824	29/05/19	Advanced Authorisation	Koninklijke Bibliotheek		Requested	   	
65536	07/05/19	Blanket Authorisation	Dummy Authorising Agency		Approved	   	

3. Go to the Permissions tab.



Harvest Authorisations

Advanced Authorisation

General | URL Patterns | Authorising Agencies | **Permissions**

Id: 589824
Agency: workshop

Title: *

Description:

Order No:

Published:

Enabled:

Annotations

Date	User	Notes	Action
There are no annotations available.			

4. Open the existing permission in edit mode.



5. Enter the following detail, and click Save.

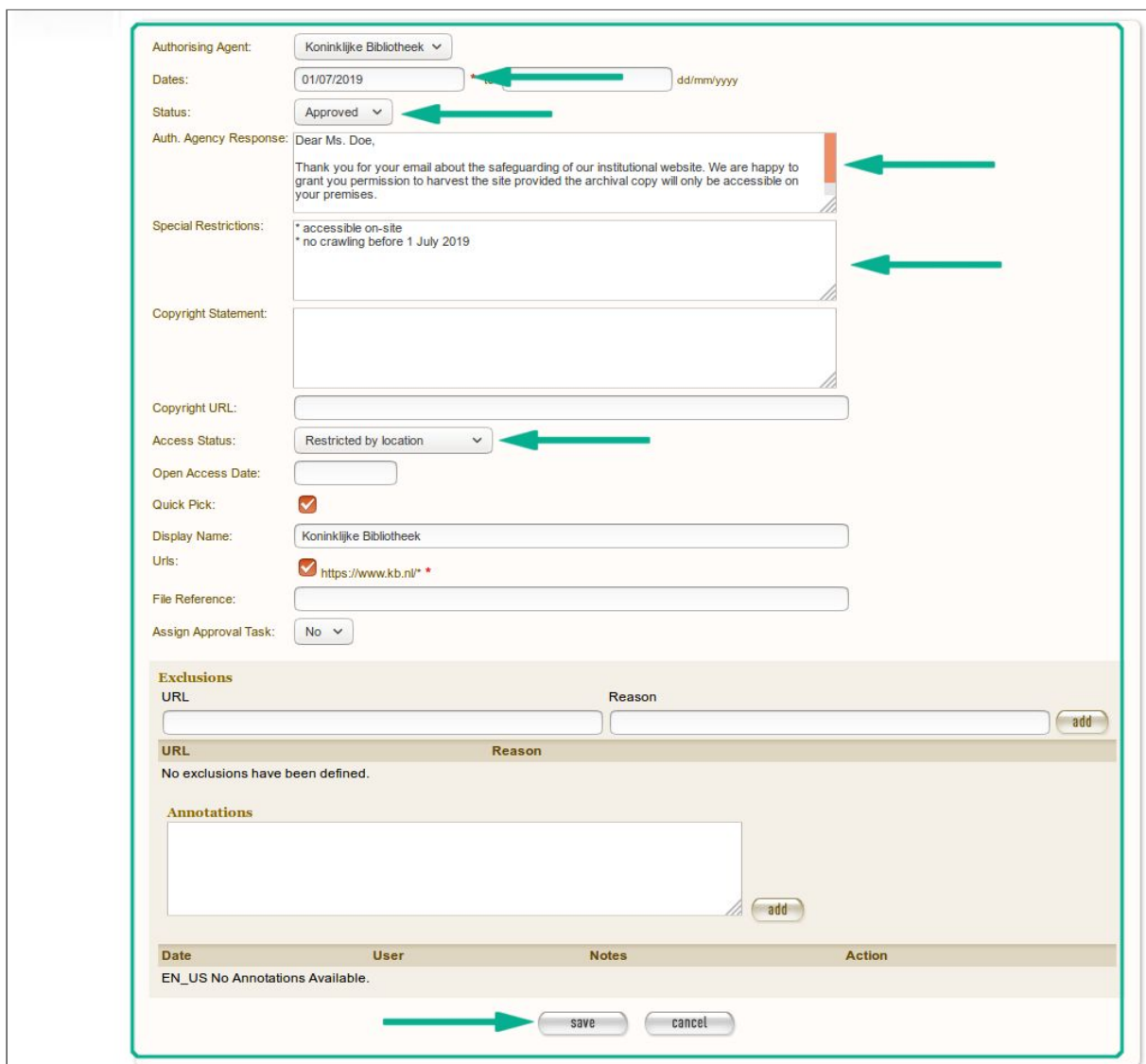
Dates: <a future date> to <blank>

Status: Approved

Auth. Agency Response: <paste the body of the email you've received from the site owner>

Special Restrictions: <note the restrictions on crawl start time and access>

Access Status: Restricted by location



6. Finally, save the Harvest Authorisation.

The screenshot shows the 'Harvest Authorisations' interface. At the top, there is a green header with a globe icon and a checkmark. Below the header, the title 'Harvest Authorisations' is displayed in green, followed by 'Advanced Authorisation'. There are four tabs: 'General', 'URL Patterns', 'Authorising Agencies', and 'Permissions'. A 'create new' button is located in the top right corner. Below the tabs is a table with the following data:

Status	Date Requested	Authorising Agent	From	To	URL Patterns	Action
Pending		Koninklijke Bibliotheek	28/05/2019		https://www.kb.nl/*	

Below the table, there are 'save' and 'cancel' buttons. A green arrow points to the 'save' button.

Attempting a crawl

Since the Harvest Authorisation has been approved, we are now able to create a Target that is governed by this Authorisation.

1. Navigate to the Targets tab.
2. Create a new Target.
3. In the General tab enter the following detail
Name: <short name of your institution>
Description: <full name of your institution> website
State: Approved

The screenshot shows the 'Targets' interface. At the top, there is a red header with a target icon. Below the header, the title 'Targets' is displayed in red, followed by 'KB'. There are eight tabs: 'General', 'Seeds', 'Profile', 'Schedule', 'Annotations', 'Description', 'Groups', and 'Access'. The 'General' tab is selected. Below the tabs is a form with the following fields:

- Id:** 720896
- Name:** KB
- Description:** Koninklijke Bibliotheek website
- Reference Number:** (empty field)
- Run on Approval:**
- Use Automated QA:**
- Owner:** Workshop User
- State:** Approved
- Auto-prune:**
- Reference Crawl:**
- Request to Archivists:** (empty text area)

At the bottom of the form, there are 'save' and 'cancel' buttons. Green arrows point to the 'Name' field, the 'Description' field, and the 'State' dropdown menu.

- In the Seeds tab, add a URL that conforms to the URL pattern that you added under the Permissions tab of the Harvest Authorisation earlier.
- From the Authorisation dropdown, select the one that corresponds to the name you set as Display Name under the Permissions tab earlier (it's listed here, because you checked Quick Pick under Permissions).
- Click on the link button. This will link the Harvest Authorisation to this seed.

- In the Schedule tab, select Harvest Now and save the Target.

- A crawl will be started in 5 minutes. Of course we are a bit impatient, so we want to start it now. To do that go to the Queue.

9. And click on the Harvest Now button of our newly created Target Instance.

Target Summary

Search

ID: From: To: 28/08/2019 23:59:59 Agency: Owner:

Name: Flag: State: Scheduled Queued Running Paused Harvested Archived Archiving Investigate Failed

QA Recommendation: Archive Reject Delist Endorsed Rejected Archived

Non-Display Only: Aborted Endorsed Rejected Archived

Multi-select Action: delist archive endorse delete reject

Thumbnail	Name	Harvest Date	State	Owner	Run Time	Data Downloaded	URLs	% Failed	Crawls	QA Recom	Action
<input type="checkbox"/>	--	KB	29/05/2019 14:26:06	Scheduled	W. User		0.0	1			<input type="button" value="archive"/>

10. Finally click on the allocate button to actually assign the crawl to a Harvest Agent.

Target Summary

KB (786433)

Harvest Now

Id: 786433
Target Name: KB
Schedule: 30/05/2019 14:26:06

Harvest Agent	Max Harvests	Current Harvests	
H3 Agent	2	0	<input type="button" value="allocate"/>

11. You should now get an error, because, according to the Permissions, we are not allowed to crawl the site until the future date that you set in the previous section.

Target Summary

The following validation errors have occurred.
The target 720896 is not approved for harvest.

Harvest Now

Id: 786433
Target Name: KB
Schedule: 30/05/2019 14:26:06

Harvest Agent	Max Harvests	Current Harvests	
H3 Agent	2	0	<input type="button" value="allocate"/>

Tutorial D - Advanced Crawling and Quality Review

What you will learn [seeds, profiles, scheduling, groups, crawl monitoring, Heritrix 3 scripting, quality review]

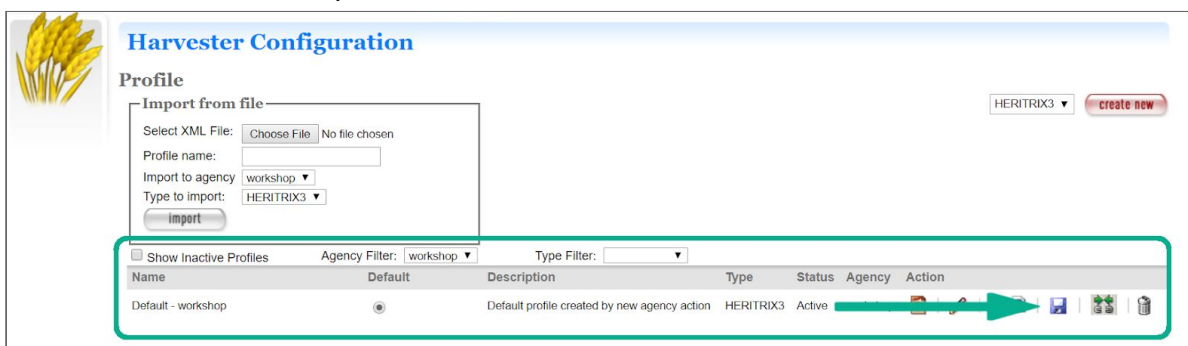
Scenario: You must crawl a website that is under a domain that your institution has authorisation to harvest. This website contains additional subdomains that also need to be captured. By default, extra unwanted content will be captured also, that must be pruned during quality review.

Importing Profiles

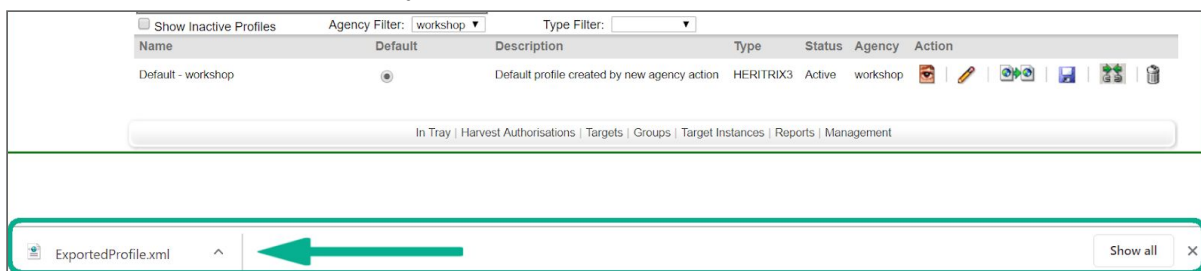
Advanced configuration of profiles can be done through WCT by importing them. This allows for additional Heritrix configuration that is beyond the basic profile options available through the WCT user interface. To make advanced configuration changes, the raw xml of a Heritrix profile has to be edited, and then imported into WCT.

For more information on WCT profiles, see the [documentation](#).

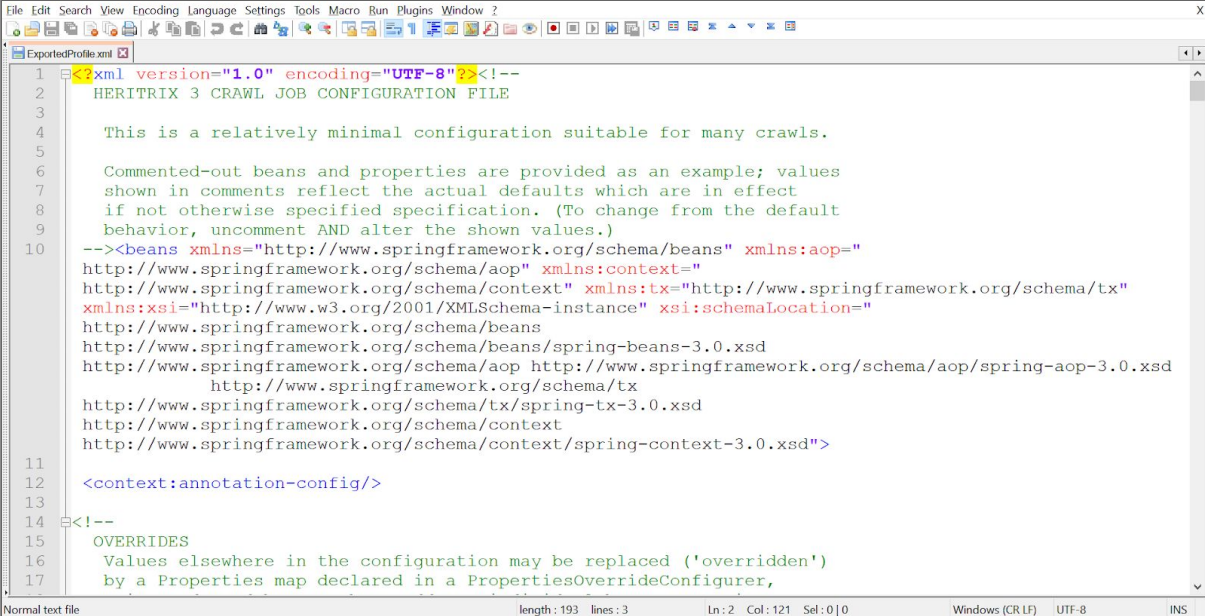
1. Navigate to the Management tab in WCT
2. Click the Profile option under Harvester Configuration.
3. You should see the default profile listed - "Default - workshop". Click the Export icon next to the default profile.



4. A download will start in your browser for an xml file.



5. Open the xml file in a suitable text editor, preferably one that supports xml syntax highlighting.



```
1 <?xml version="1.0" encoding="UTF-8"?><!--
2 HERITRIX 3 CRAWL JOB CONFIGURATION FILE
3
4 This is a relatively minimal configuration suitable for many crawls.
5
6 Commented-out beans and properties are provided as an example; values
7 shown in comments reflect the actual defaults which are in effect
8 if not otherwise specified specification. (To change from the default
9 behavior, uncomment AND alter the shown values.)
10 --><beans xmlns="http://www.springframework.org/schema/beans" xmlns:aop="
http://www.springframework.org/schema/aop" xmlns:context="
http://www.springframework.org/schema/context" xmlns:tx="http://www.springframework.org/schema/tx"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="
http://www.springframework.org/schema/beans
http://www.springframework.org/schema/beans/spring-beans-3.0.xsd
http://www.springframework.org/schema/aop http://www.springframework.org/schema/aop/spring-aop-3.0.xsd
http://www.springframework.org/schema/tx
http://www.springframework.org/schema/tx/spring-tx-3.0.xsd
http://www.springframework.org/schema/context
http://www.springframework.org/schema/context/spring-context-3.0.xsd">
11 <context:annotation-config/>
12
13
14 <!--
15 OVERRIDES
16 Values elsewhere in the configuration may be replaced ('overridden')
17 by a Properties map declared in a PropertiesOverrideConfigurer,
```

6. Choose one or more of the following examples and update the profile xml. Or make your own configuration change if you have knowledge of working with Heritrix profiles.

- a. Increase the maximum repetitions allowed in URL paths. Remove the xml comment syntax (<!--, -->), and increase the value to 10.

```
<bean class="org.archive.modules.deciderules.PathologicalPathDecideRule">
  <!-- <property name="maxRepetitions" value="2" /> -->
</bean>
```

- b. Reduce the amount of time Heritrix waits before giving up on a failing URL. Remove the xml comment syntax (<!--, -->), and reduce the *retryDelaySeconds* value to 600 and the *maxRetries* value to 10.

```
<bean class="org.archive.crawler.frontier.BdbFrontier" id="frontier">
  <!-- <property name="snoozeLongMs" value="300000" /> -->
  <!-- <property name="retryDelaySeconds" value="900" /> -->
  <!-- <property name="maxRetries" value="30" /> -->
```

- c. Configure a http proxy for Heritrix to crawl through, and its credentials. Remove the xml comment syntax (<!--, -->), and enter the appropriate proxy configuration.

```
<bean class="org.archive.modules.fetcher.FetchHTTP" id="fetchHttp">
  <!-- <property name="httpProxyHost" value="" /> -->
  <!-- <property name="httpProxyPort" value="0" /> -->
  <!-- <property name="httpProxyUser" value="" /> -->
  <!-- <property name="httpProxyPassword" value="" /> -->
```

- Once you have edited and saved the profile xml, import it back into WCT, using the following details.

Select XML File: <browse and select the profile you just edited>

Profile name: Special profile

Import to agency: workshop

Type to import: HERITRIX3

Name	Default	Description	Type	Status	Agency	Action
Default - workshop	<input checked="" type="radio"/>	Default profile created by new agency action	HERITRIX3	Active	workshop	[edit] [delete] [refresh] [help] [info] [trash]

- You should see the new imported profile listed - "Special Profile". Click the edit icon next to it.

Name	Default	Description	Type	Status	Agency	Action
Default - workshop	<input checked="" type="radio"/>	Default profile created by new agency action	HERITRIX3	Active	workshop	[edit] [delete] [refresh] [help] [info] [trash]
Special Profile	<input type="radio"/>	Imported profile	HERITRIX3	Active	workshop	[edit] [delete] [refresh] [help] [info] [trash]

- Navigate to the Scope tab.

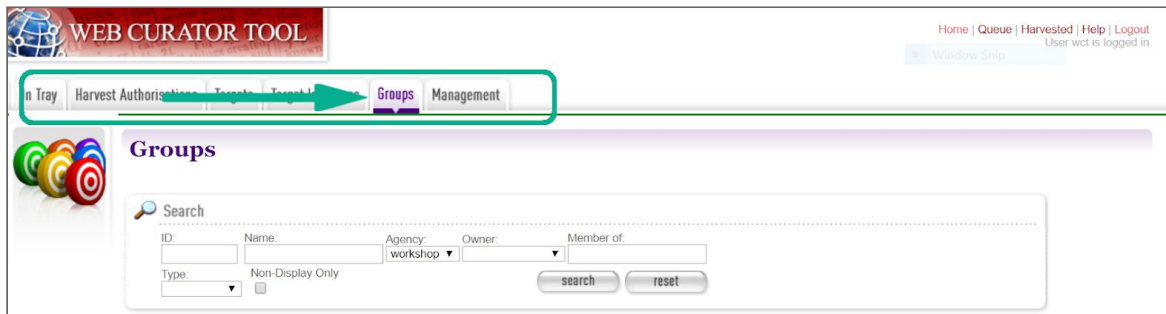
When editing/viewing an imported profile you will see an XML editor instead of the standard WCT Profile UI. This is done to remove the risk of WCT misreading a profile edited by a user, as the validity of the xml cannot be guaranteed.

Groups

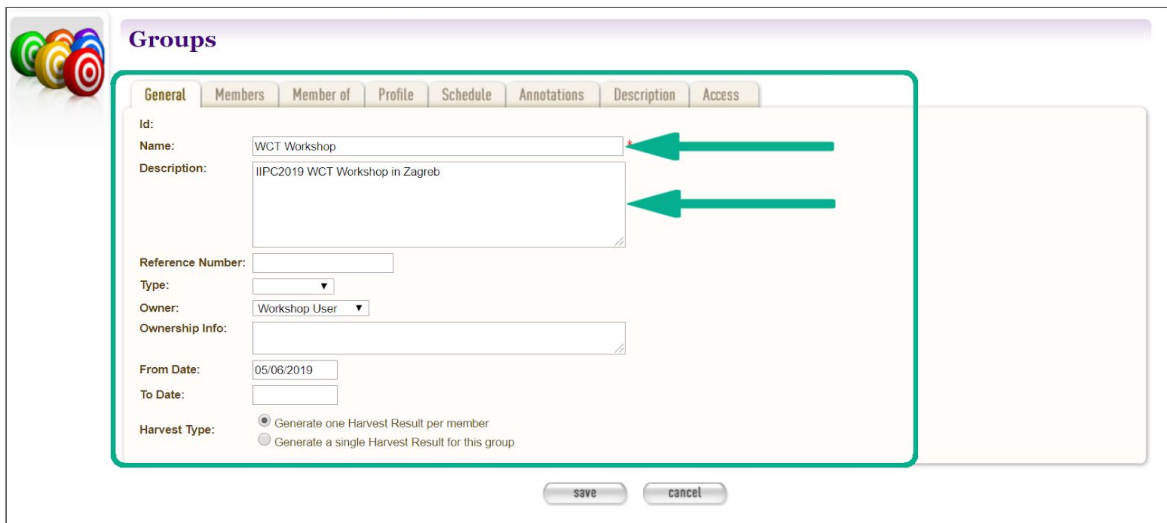
Groups are a way of associating Targets within WCT. For instance, this provides a way to group collections of websites based on themes, subjects and events. Nested groups are also possible.

For more information on Groups in WCT, see the [documentation](#).

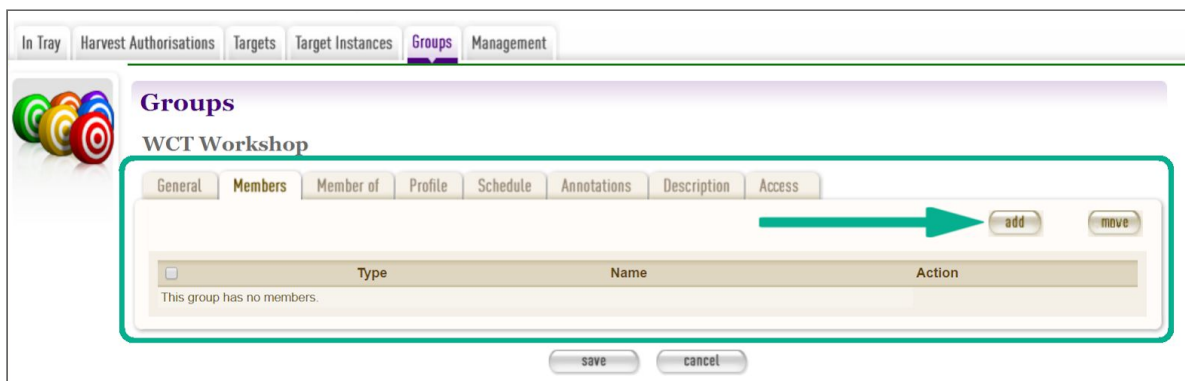
1. Navigate to the Groups tab within WCT.



2. Click the Create New button
3. In the General tab, enter the following detail
Name: WCT Workshop
Description: IIPC2019 WCT Workshop in Zagreb



4. Navigate to the Members tab, and click the Add button.



5. Tick the checkboxes next to all the Targets you have created during this workshop.
Click Save to assign them to this group.

Selected				
No Selection				
Results				
	Name	Type	Status	Agency
<input checked="" type="checkbox"/>	NLNZ	Target	Completed	workshop
<input checked="" type="checkbox"/>	IIPC	Target	Approved	workshop

Setup Target

A Target can contain one or more seed URLs for a crawl. This is useful if you know that the general path of the Heritrix crawler will not capture all the content of the website you want to harvest. For instance if the website has various sub-domains or you are collecting content across multiple domains/websites.

We will use the same website as Tutorial B, but slightly modify the name of the Target.

For more information on WCT Targets, see the [documentation](#).

1. Navigate to the Targets tab in WCT.
2. Create a new Target.
3. In General tab, enter the details for a new website you want to harvest.
Name: <short name of your institution> + “ - Tutorial D”
Description: <full name of your institution> website
State: Approved

Targets

General

Seeds

Profile

Schedule

Annotations

Description

Groups

Access

Id:

Name:

Description:

National Library of New Zealand website

Reference Number:

Run on Approval:

Use Automated QA:

Owner:

State:

Auto-prune:

Reference Crawl:

Request to Archivists:

- Navigate to the Seeds tab, and link two or more seed URLs using the default Harvest Authorisation. Ensure the correct URL is marked as the primary seed.

Targets
NLNZ - Tutorial D

General Seeds Profile Schedule Annotations Description Groups Access

Seed:
 Authorisation: Auto

Seed	Primary	Harvest Auth	Auth Agent	Start	End	Status	Action
<input type="checkbox"/> https://digitalpreservation.natlib.govt.nz/	<input type="checkbox"/>	Blanket Authorisation	Dummy Authorising Agency	01/05/2019	30/04/2029	Approved	<input type="button" value="add"/> <input type="button" value="edit"/> <input type="button" value="delete"/>
<input type="checkbox"/> http://mp.natlib.govt.nz/	<input type="checkbox"/>	Blanket Authorisation	Dummy Authorising Agency	01/05/2019	30/04/2029	Approved	<input type="button" value="add"/> <input type="button" value="edit"/> <input type="button" value="delete"/>
<input type="checkbox"/> https://natlib.govt.nz/	<input checked="" type="checkbox"/>	Blanket Authorisation	Dummy Authorising Agency	01/05/2019	30/04/2029	Approved	<input type="button" value="add"/> <input type="button" value="edit"/> <input type="button" value="delete"/>

- Navigate to the Profile tab, and select the new imported profile "Special Profile". Ticking the *Override Imported Profile* checkbox will display the xml editor for making Target specific changes to the profile.

Targets
NLNZ - Tutorial D

General Seeds Profile Schedule Annotations Description Groups Access

Base Profile
 Harvester Type: HERITRIX3
 Base Profile: Special Profile
 Profile Overrides
 Override Imported Profile:
 Profile Note:

- Navigate to the Groups tab, and click the Add button.

Targets
NLNZ - Tutorial D

General Seeds Profile Schedule Annotations Description Groups Access

Name	Action
This target does not belong to any groups.	

7. Tick the checkbox next to the *WCT Workshop* group, and click Save.

8. Navigate to the Schedule tab, tick the Harvest Now box, and click Save.
9. Click the View Target Instances icon next to the Target.
10. Start the Target Instance using the Harvest Now option.

Heritrix 3 Scripting Console

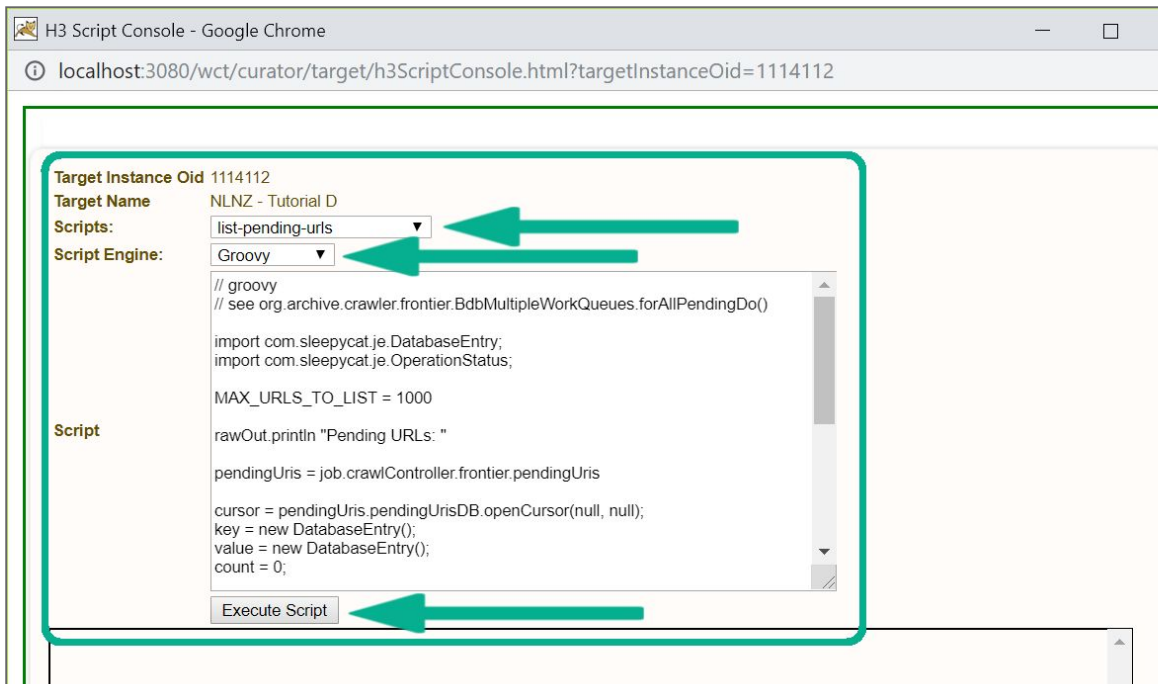
Running Heritrix 3 Target Instances have an H3 script console available to use. This console can be used to run scripts against the Target Instance job in Heritrix 3, similar to the scripting console available in H3's own user interface.

1. Observe the scheduled Target Instance move into a Running state. Click the H3 Script Console icon.

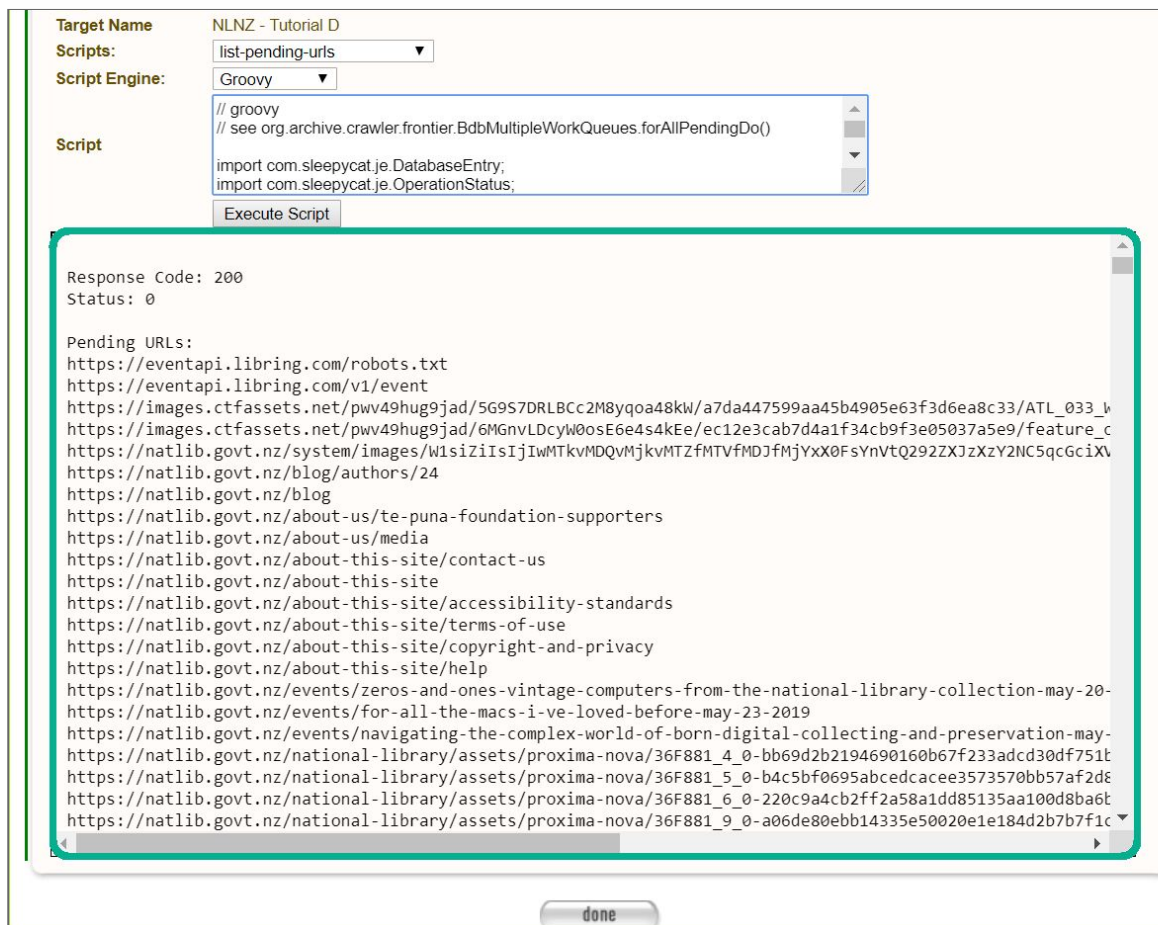
2. Here we can edit and execute predefined and custom scripts against the running crawl. Enter the following details and click the Execute Script button.

Scripts: list-pending-urls

Script Engine: Groovy



3. The *list-pending-urls* script will return the next 1000 URLs from the crawl queue.



4. Feel free to experiment with the other two available scripts.
 - The *empty-frontier* script will remove all remaining URLs from the crawl queue.
 - The *remove-host-from-frontier* script will remove all URLs containing a specified hostname from the crawl queue. That hostname is set in the script.
5. Further scripts can be found on the [Heritrix 3 wiki](#).

Quality review

During the quality review of a Target Instance, it is typical practice to identify URLs that are missing from a harvest, or undesirable URLs that were crawled. Depending on the scale, these issues can sometimes be fixed on a harvested Target Instance rather than harvesting a new one.

1. Navigate back to the Target Instance tab.
2. Stop the running Target Instance.
3. Once it has finished and moved into the Harvested state, click the Edit icon.
4. Navigate to the Harvest Results tab and wait for the indexing to complete.

Target Summary
NLNZ - Tutorial D (1114112)

General Profile Harvest State Logs **Harvest Results** Annotations Display

No.	Date	Derived From	User	Notes	State	Action
1	03/06/2019 10:15:16		W User	Original Harvest	Indexing	Restart Indexing

save cancel

5. Once the Harvest Result has been indexed, click the Review link.
6. Notice the multiple seed URLs listed. These can each be viewed directly in OpenWayback by clicking the *Review in Access Tool* link.

Quality Review Tools

Browse

https://natlib.govt.nz/	Review this Harvest Review in Access Tool Live Site Archives Harvested Web Archive
https://digitalpreservation.natlib.govt.nz/	Review this Harvest Review in Access Tool Live Site Archives Harvested Web Archive
http://mp.natlib.govt.nz/	Review this Harvest Review in Access Tool Live Site Archives Harvested Web Archive

Tool	Description
Harvest History	Compare current harvest result with previous harvests.
Tree View	Graphical view of harvested data.

done

7. Click the *Tree View* link

Quality Review Tools

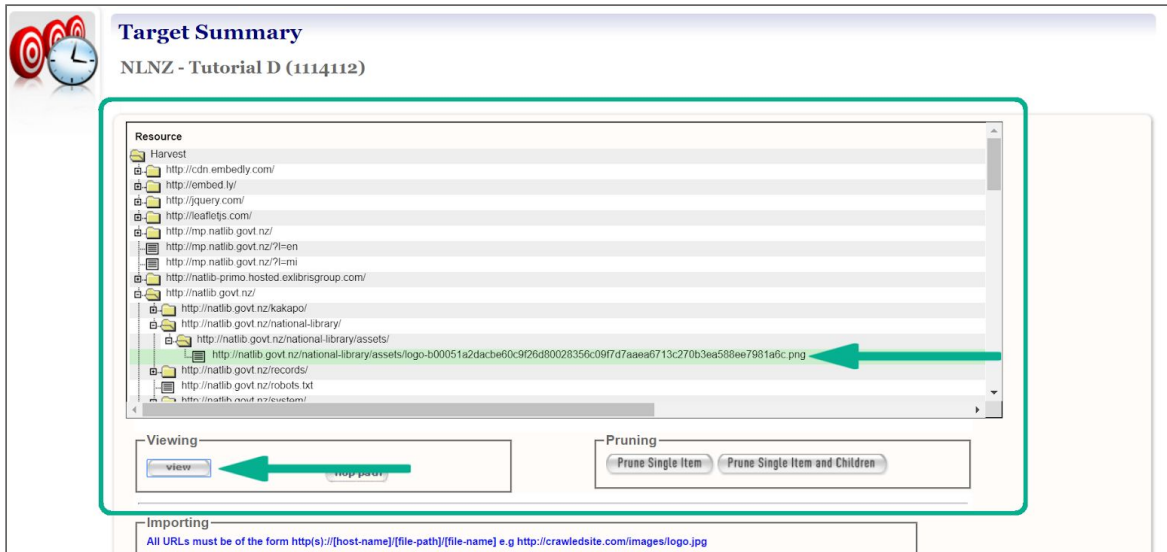
Browse

https://natlib.govt.nz/	Review this Harvest Review in Access Tool Live Site Archives Harvested Web Archive
https://digitalpreservation.natlib.govt.nz/	Review this Harvest Review in Access Tool Live Site Archives Harvested Web Archive
http://mp.natlib.govt.nz/	Review this Harvest Review in Access Tool Live Site Archives Harvested Web Archive

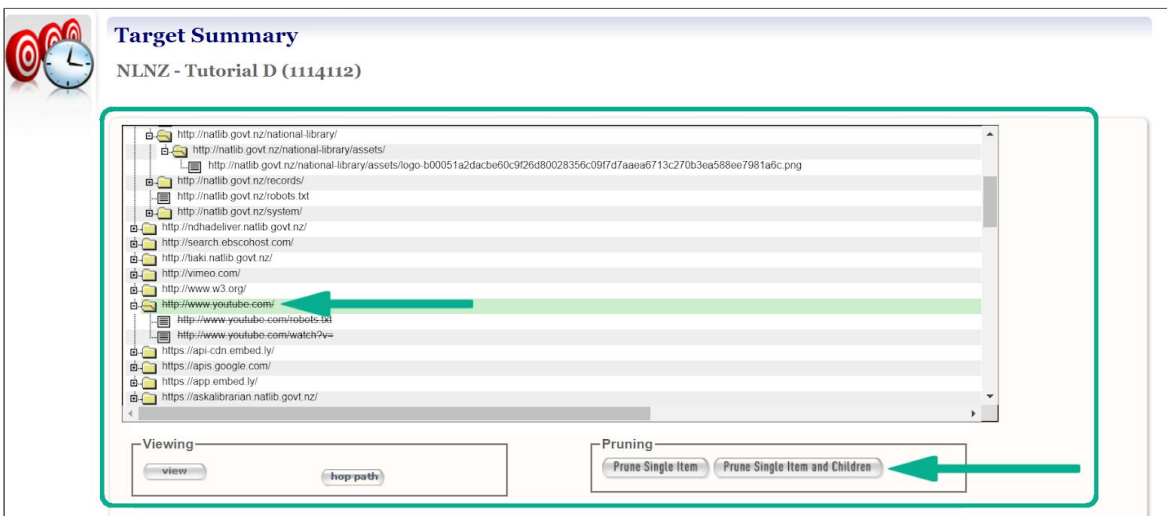
Tool	Description
Harvest History	Compare current harvest result with previous harvests.
Tree View	Graphical view of harvested data.

done

8. All the harvested resources can be viewed within a tree structure. Highlight an individual resource and click the *View* button to show the document directly using OpenWayback.



9. The tree view can be used to prune unwanted URLs from a particular hostname or sub-path. Highlight a folder in the tree view, and click the *Prune Single Item and Children* button.



10. The resources with the ~~strike through~~ are now staged for pruning. Enter the following details and click Save.

Provenance Note: Removing unwanted content from crawl.

The screenshot shows a web interface for importing content. It has three main sections: 'Import from disk file', 'Import from URL', and 'Automated QA Imports'. The 'Automated QA Imports' section shows 'There are no URLs available for import.' Below this is a 'Provenance Note' section with a text area containing 'Removing unwanted content from crawl' and a 'save' button. A green box highlights the 'Provenance Note' section, and two green arrows point to the text area and the 'save' button.

11. A second Harvest Result is now being indexed, containing the modifications just made through pruning. Once indexed, this will also be available for review.

The screenshot shows the 'Target Summary' page for 'NLNZ - Tutorial D (1114112)'. The 'Harvest Results' tab is active. Below the tabs is a table with the following data:

No.	Date	Derived From	User	Notes	State	Action
1	03/06/2019 10:15:16		W. User	Original Harvest	Review	Review Endorse Reject for Reason: No new content
2	03/06/2019 10:40:56	1	W. User	Removing unwanted content from crawl	Indexing	Restart Indexing

The second row is highlighted with a green box. Below the table are 'save' and 'cancel' buttons.

12. You can decide which Harvest Results to endorse (and ultimately archive), or reject.

The screenshot shows the 'Target Summary' page for 'NLNZ - Tutorial D (1114112)'. The 'Harvest Results' tab is active. Below the tabs is a table with the following data:

No.	Date	Derived From	User	Notes	State	Action
1	03/06/2019 10:15:16		W. User	Original Harvest	Rejected	Rejection Reason: Technical reasons
2	03/06/2019 10:40:56	1	W. User	Removing unwanted content from crawl	Endorsed	Submit to Archive UnEndorse

The second row is highlighted with a green box. Below the table are 'save' and 'cancel' buttons.

Tutorial E - User Management

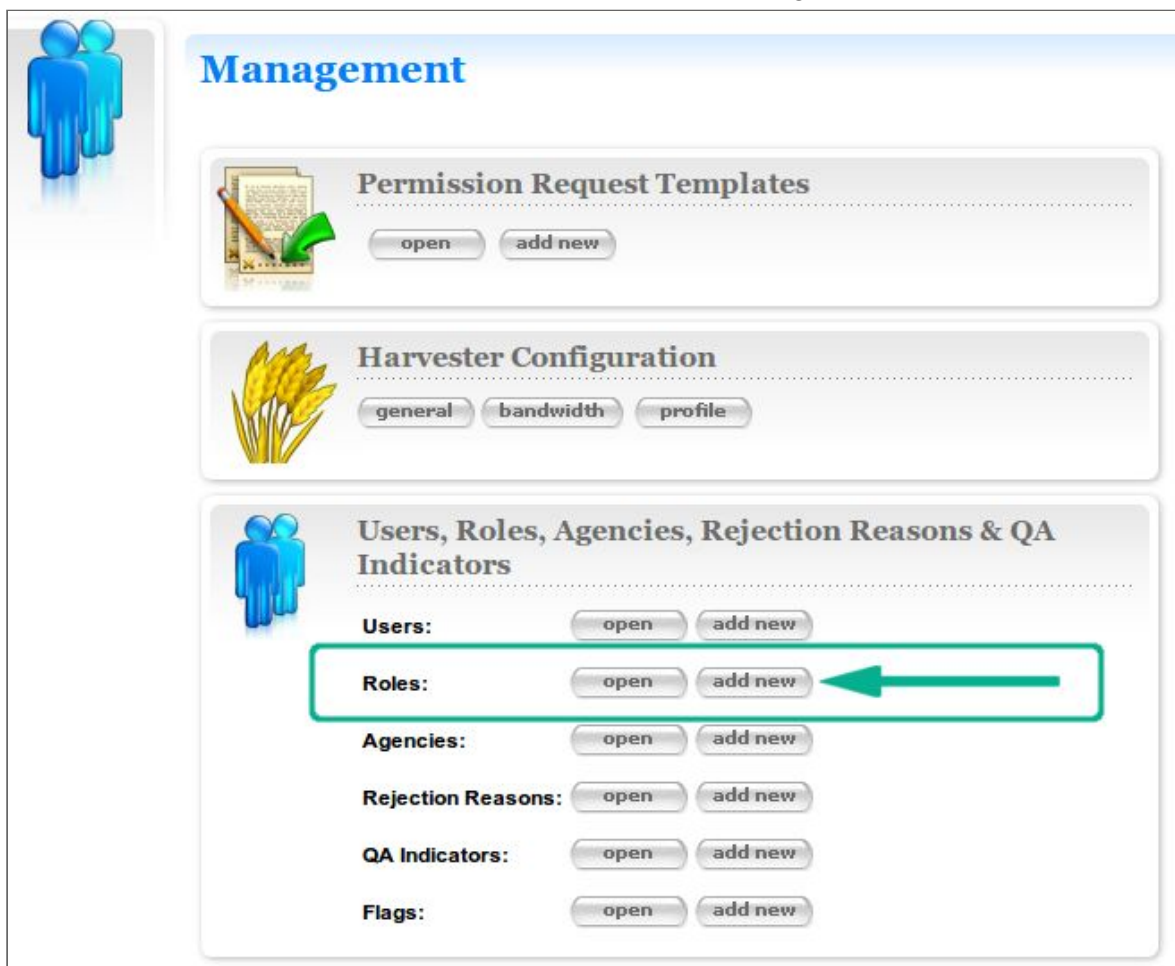
What you will learn [user management]

Scenario: You have to setup a new user, who will have restricted access to WCT.

In this tutorial we will create a new user with restricted permissions who will be assigned specific tasks within the WCT workflow. Let's suppose someone in your organisation is tasked with quality control and endorsement of completed harvests. We will create a user with the corresponding privileges.

Creating the role

1. Click on the add new button behind Roles in the Management tab.



The screenshot shows the 'Management' interface. On the left is a sidebar with a blue icon of two people. The main content area is titled 'Management' and contains three sections:

- Permission Request Templates**: Includes an icon of a document and pencil, and buttons for 'open' and 'add new'.
- Harvester Configuration**: Includes an icon of wheat stalks and buttons for 'general', 'bandwidth', and 'profile'.
- Users, Roles, Agencies, Rejection Reasons & QA Indicators**: Includes an icon of two people and a list of categories, each with 'open' and 'add new' buttons:
 - Users: open, add new
 - Roles: open, add new** (highlighted with a green box and a green arrow pointing to the 'add new' button)
 - Agencies: open, add new
 - Rejection Reasons: open, add new
 - QA Indicators: open, add new
 - Flags: open, add new

2. Select your agency.
3. In Role Name, enter: 'QA'.
4. In Description, enter: 'QA role'
5. Select the roles: Login, Update User Credentials.

Users, Roles, Agencies & Rejection Reasons

Role

Agency: workshop

Role Name: QA

Description: QA role

Select All

Login

Login

Update User Credentials

N/A

Owner

Manage Copying Permissions and Access Rights

Create Harvest Authorisations

Modify Harvest Authorisations

Confirm Permissions

Agency

All

Agency

6. Select all the roles under *Manage Harvests*.

Set Harvest Profile Level 1

Set Harvest Profile Level 2

Set Harvest Profile Level 3

Agency

Agency

Agency

Manage Harvests

Manage Target Instances

Launch Target Instance Immediate

Manage Web Harvester System

Endorse Harvest

UnEndorse Harvest

Archive Harvest

Agency

Agency

All

Agency

Agency

Agency

Manage Target Groups

Create Target Group

Add Target to Group

Agency

Owner

7. Save the role.

Ownership

Give Ownership to Agency ▾

Take Ownership from Agency ▾

save cancel

Create the user

Now, let's create a user with this role.

1. Click on the add new button behind Users in the Management tab.

Management

Permission Request Templates
open add new

Harvester Configuration
general bandwidth profile

Users, Roles, Agencies, Rejection Reasons & QA Indicators

Users: open add new

Roles: open add new

Agencies: open add new

Rejection Reasons: open add new

QA Indicators: open add new

Flags: open add new

2. Use your own name in First Name, Last Name and Username.

3. Enter and confirm a password.

This is a temporary password. The first time a user logs in, they are forced to set their own new password.

4. Select your agency.

5. In the Email field, just put: jane.doe@test.email.address.

6. Save the user.

Users, Roles, Agencies & Rejection Reasons

User

Credentials

Title: e.g. Mr, Mrs, Ms

First Name: * Last Name: *

Username: *

Use External Directory for Password Credentials

Password: * Confirm Password: *

Contact Information

Agency: *

Address:

Phone:

Email: *

Receive notifications for Harvester Warnings

Receive general notifications

Receive notifications by email

Receive tasks by email

7. Open the group memberships screen for the new user.

Users, Roles, Agencies & Rejection Reasons

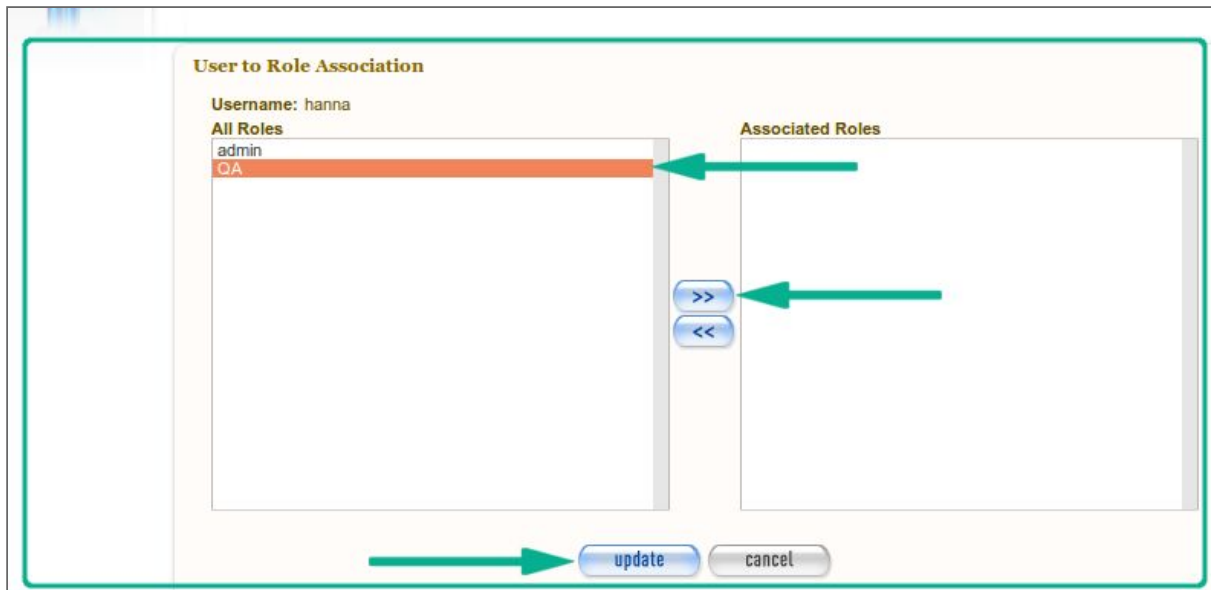
The user 'hanna' has been created successfully.

Users Agency Filter:

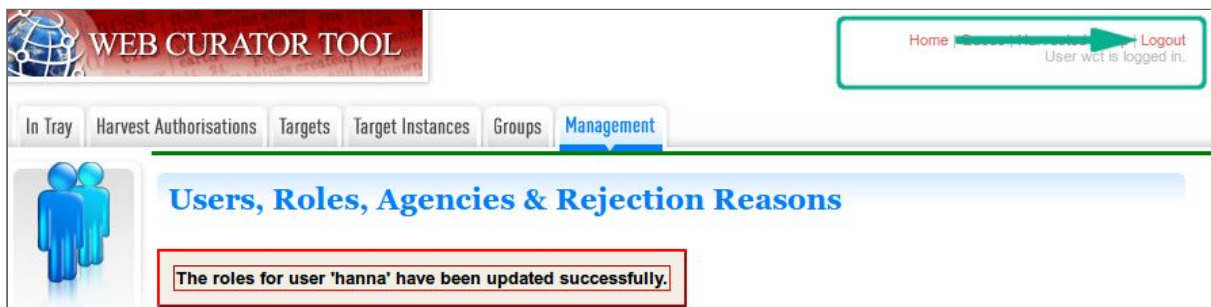
Username	First name	Last name	Email	Agency	Action
bootstrap	System	Bootstrap	bootstrap@sytec.co.nz	bootstrap	
hanna	Hanna	Koppelaar	jane.doe@test.email.address	workshop	
test	test	test	jane.doe@test.email.address	workshop	

8. Select the role 'QA' and assign it to this user.

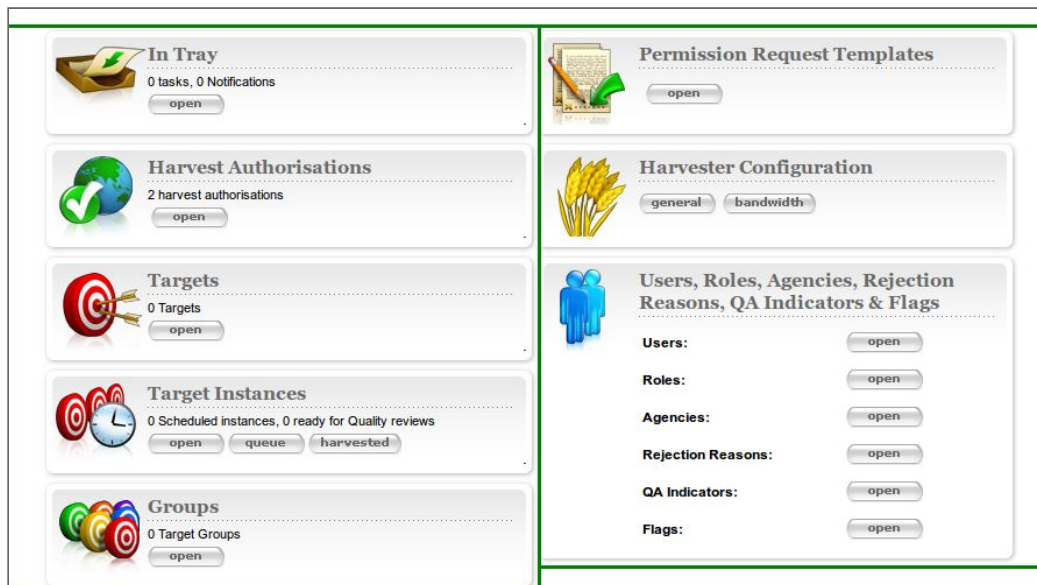
9. Click on the update button.



10. Logout.



11. Now login as the new user you've just created. You will be asked to change your password, since this is your first login. Notice that the interface now offers less options, because this user has less privileges. E.g. when you open the Targets screen, you will not see the button that allows you to create a new Target.

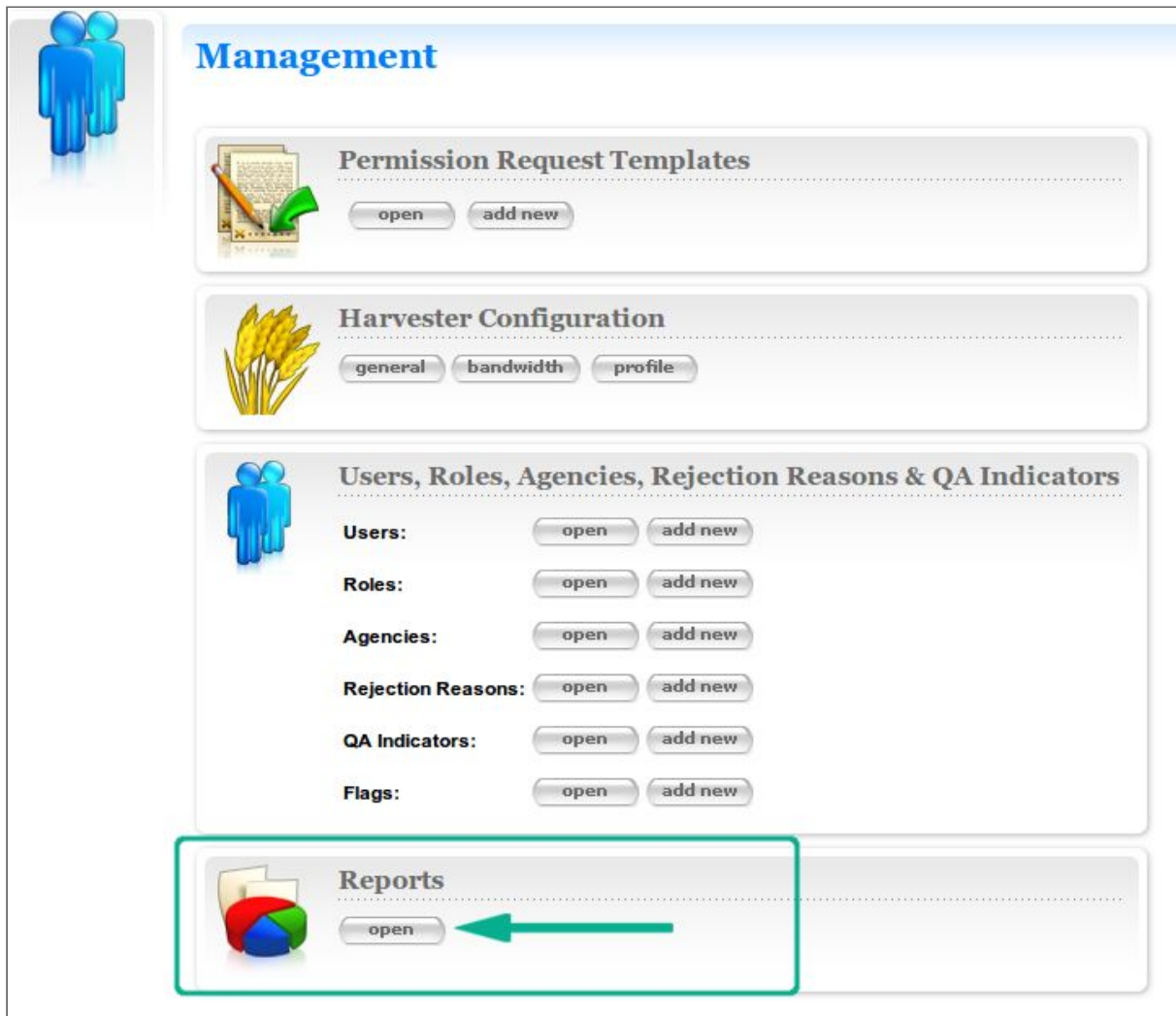


Tutorial F - Reporting

What you will learn [reporting]

Scenario: You have been asked to provide a summary of the past month's crawling activity. Run some reports.

1. Click on the open button under Reports in the Management tab.



Management

Permission Request Templates
open add new

Harvester Configuration
general bandwidth profile

Users, Roles, Agencies, Rejection Reasons & QA Indicators

Users: open add new

Roles: open add new

Agencies: open add new

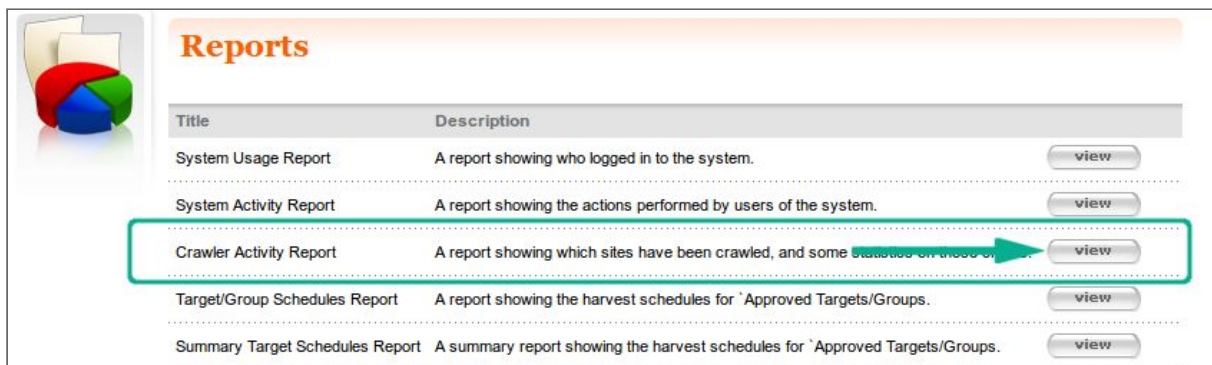
Rejection Reasons: open add new

QA Indicators: open add new

Flags: open add new

Reports
open

2. Click on the view button behind the Crawler Activity Report.



Reports

Title	Description	
System Usage Report	A report showing who logged in to the system.	view
System Activity Report	A report showing the actions performed by users of the system.	view
Crawler Activity Report	A report showing which sites have been crawled, and some statistics on those sites.	view
Target/Group Schedules Report	A report showing the harvest schedules for 'Approved Targets/Groups.	view
Summary Target Schedules Report	A summary report showing the harvest schedules for 'Approved Targets/Groups.	view

- In Start Date enter the first date of last month, in End Date enter the date of tomorrow.
- Generate the report.

Reports

Crawler Activity Report
A report showing which sites have been crawled, and some statistics on those crawls.

*Start Date is inclusive.
End Date is exclusive*

Start Date: * dd/MM/yyyy

End Date: * dd/MM/yyyy

Agencies: (Optional)

Users: (Optional)

- You can download ('save'), print or email the report in either HTML or CSV format. *Note that email will not work if the server on which your WCT is running does not have SMTP enabled.*

Reports

Crawler Activity Report

2 results:

Id	Target Name	Status	Start Date	End Date	Crawl Duration	Bytes Downloaded	Harvest Agent
458758	IIPC	Harvested	27/05/2019	27/05/2019	1296598	13605486	H3 Agent
819200	IIPC	Harvested	31/05/2019	31/05/2019	391803	109949	H3 Agent