

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.

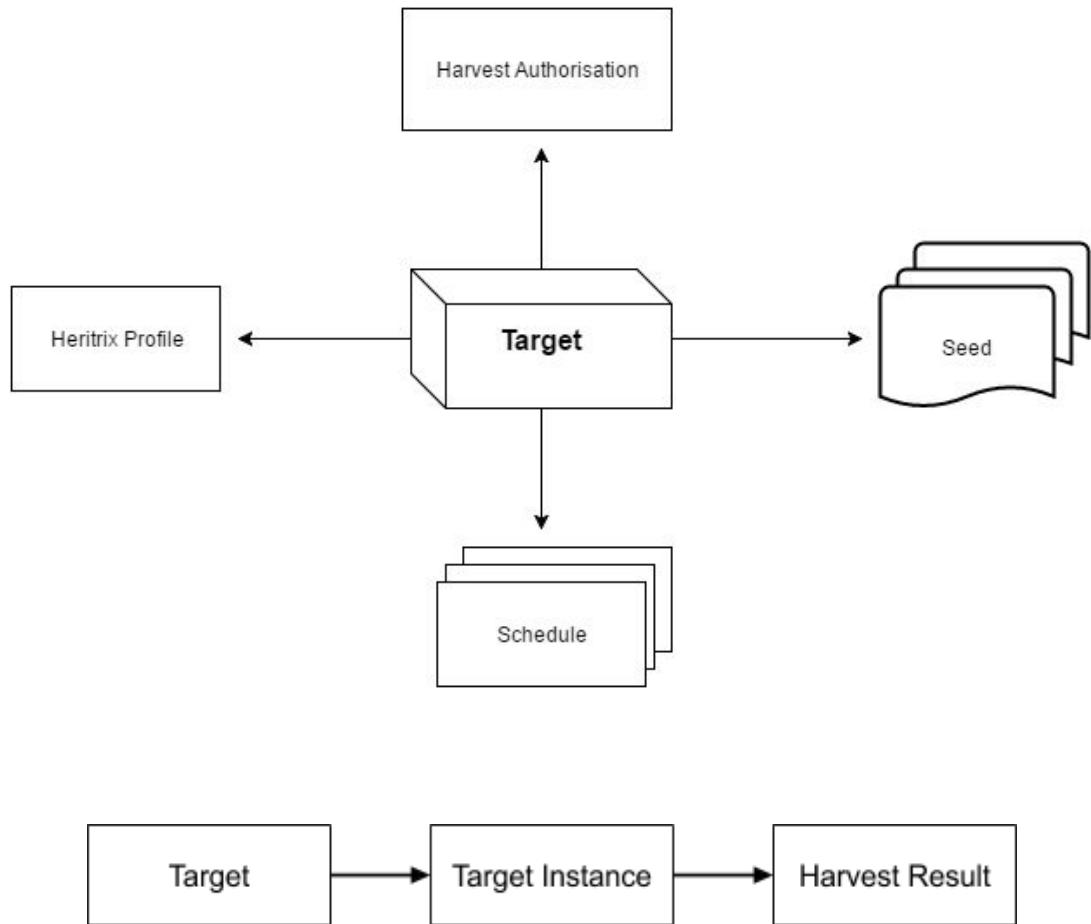
This lesson plan was written by the National Library of New Zealand and National Library of the Netherlands for the WCT Workshop held at the 2019 IIPC General Assembly, based on version 2.x of WCT.

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 it's 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

In Tray Harvest Authorisations Targets Target Instances Groups Management

Harvest Authorisations

Search

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

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

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

In Tray Harvest Authorisations Targets Target Instances Groups Management

Harvest Authorisations

General URL Patterns Authorising Agencies Permissions

Agency: workshop
Title: netpreserve.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/*`

URL Pattern Action

`http://netpreserve.org/*`

Save cancel

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

WEB CURATOR TOOL

In Tray Harvest Authorisations Targets Target Instances Groups Management

Harvest Authorisations

netpreserve.org

Authorising Agencies Permissions

create new

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

save cancel

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

WEB CURATOR TOOL

In Tray Harvest Authorisations Targets Target Instances Groups Management

Harvest Authorisations

netpreserve.org

Name: IIPC *

Description: International Internet Preservation Consortium

Contact: Jane Doe *

Phone:

Email: jane.doe@test.email.address *

Address:

save cancel

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

The screenshot shows the 'Harvest Authorisations' page for 'netpreserve.org'. The 'Permissions' tab is highlighted. A green arrow points to the 'create new' button at the top right of the form area. The form includes fields for Status, Date Requested, Authorising Agent, From, To, URL Patterns, and Action, along with '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 'Harvest Authorisations' page for 'netpreserve.org'. The 'Permissions' tab is selected. Several fields are highlighted with green arrows pointing to them:

- Dates:** A date range from '05/06/2019' to an empty field.
- Status:** Set to 'Approved'.
- Auth. Agency Response:** Text area containing 'Hi Institution, we give permission for you to harvest our website.'
- Special Restrictions:** Text area containing 'Please do not crawl our website during the weekends.'
- Urls:** Two checkboxes are checked: 'http://netpreserve.org/*' and 'https://netpreserve.org/*'.
- Copyright Statement:** An empty text area.

At the bottom, there is an 'Exclusions' section with a table for adding URL exclusions.

9. Save the new Harvest Authorisation.

The screenshot shows the WCT interface with the 'Harvest Authorisations' tab selected. A green arrow points from the 'Permissions' section to the 'save' button. The table in the 'Permissions' section contains one row:

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

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

The screenshot shows the WCT interface with the 'Management' tab selected. A green arrow points to the 'Management' tab. The 'Management' section contains two main sections: 'Permission Request Templates' and 'Harvester Configuration'.

2. Click the Profile option under Harvester Configuration.

The screenshot shows the 'Management' interface with a sidebar icon of two blue stylized figures. Below the sidebar, there's a section titled 'Permission Request Templates' with 'open' and 'add new' buttons. The main content area is titled 'Harvester Configuration' with a wheat icon. It has tabs for 'general', 'bandwidth', and 'profile'. A green arrow points to the 'profile' tab. Below this, there's another section titled 'Users, Roles, Agencies, Rejection Reasons & QA Indicators' with a 'Users:' label and 'open' and 'add new' buttons. The entire 'Harvester Configuration' section is highlighted with a green border.

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

The screenshot shows the 'Harvester Configuration' page for the 'Default - workshop' profile. The top part has fields for importing from XML and creating new profiles. Below is a table listing profiles, with 'Default - workshop' selected. The table columns are Name, Default, Description, Type, Status, Agency, and Action. The 'Action' column contains icons for edit, delete, and other actions. A green arrow points to the edit icon in this column. The bottom of the screen shows a navigation bar with links like In Tray, Harvest Authorisations, Targets, Groups, Target Instances, Reports, and Management.

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

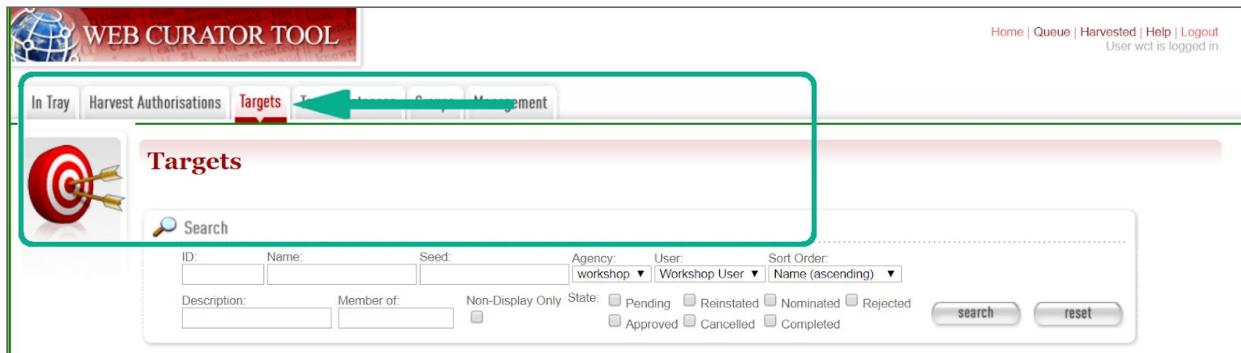
The screenshot shows the 'Harvester HERITRIX3 Configuration' page for the 'Default - workshop' profile. The 'Scope' tab is selected, indicated by a green arrow. The configuration options include Contact URL, User Agent Prefix, and various limit settings. The 'Document Limit' is set to 0, 'Data Limit' is set to 20 MB (with a green arrow pointing to it), and 'Time Limit' is set to 10 MINUTE (with a green arrow pointing to it). Other settings like Max Path Depth, Max Hops, and Extract Javascript are also visible. The bottom of the page shows an 'Include URLs' field with a regex pattern: '^rss.*|google.com/jsept.*|maps.google.*|youtube.com/embed/.*'.

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.

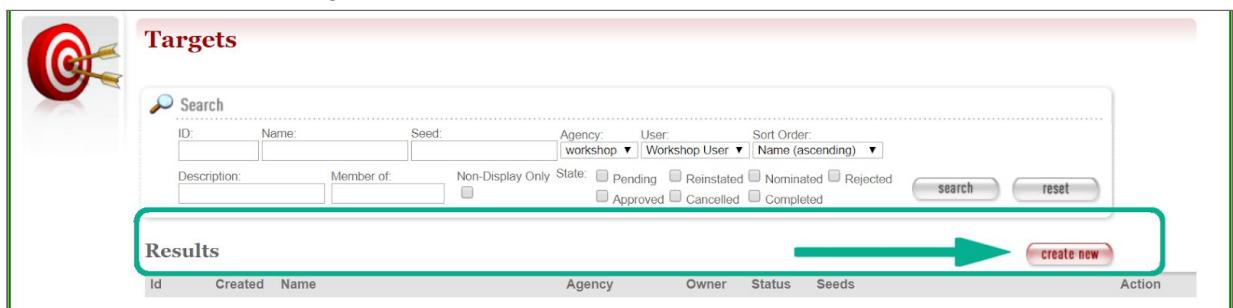
For more information on WCT Targets, see the [documentation](#).

1. Navigate to the Targets tab in WCT



The screenshot shows the 'WEB CURATOR TOOL' interface. The top navigation bar includes links for Home, Queue, Harvested, Help, and Logout. A message indicates 'User wct is logged in.' Below the navigation is a menu bar with tabs: In Tray, Harvest Authorisations, Targets (which is highlighted in red), Schedule, Annotations, Description, Groups, and Access. The main content area is titled 'Targets' and features a search bar with fields for ID, Name, Seed, Agency, User, Sort Order, Description, Member of, Non-Display Only, and State (Pending, Reinstanted, Nominated, Rejected, Approved, Cancelled, Completed). There are also 'search' and 'reset' buttons. A green arrow points to the 'Targets' tab in the menu bar.

2. Create a new Target



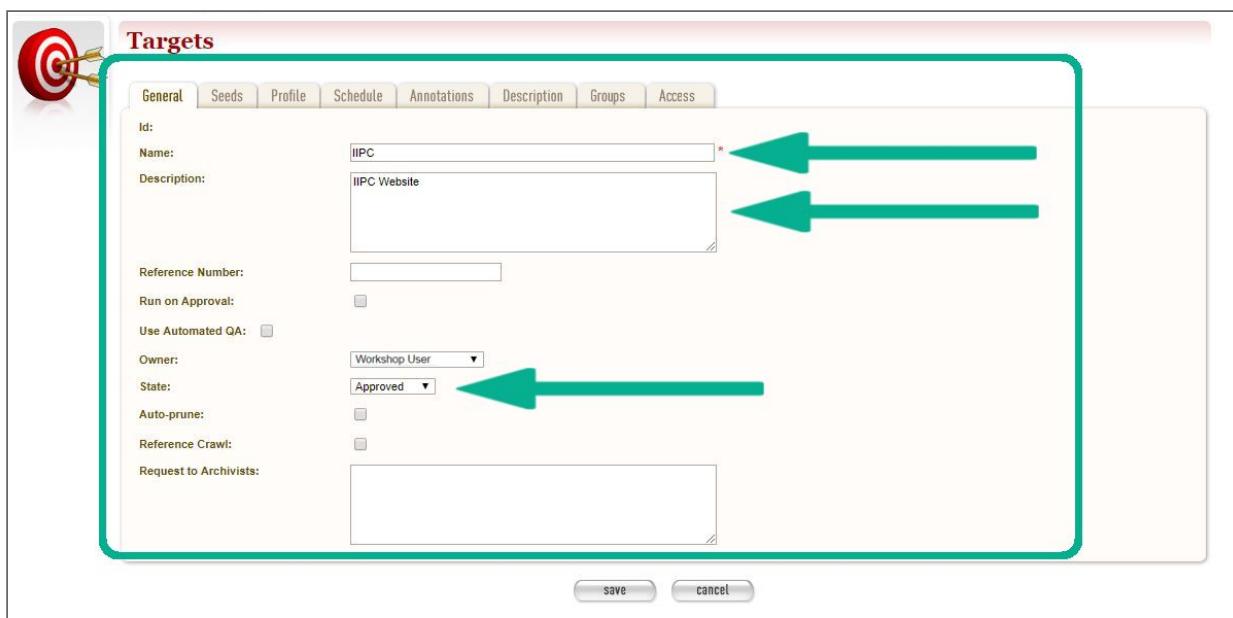
The screenshot shows the 'Targets' page. At the top is a search bar with fields for ID, Name, Seed, Agency, User, Sort Order, Description, Member of, Non-Display Only, and State (Pending, Reinstanted, Nominated, Rejected, Approved, Cancelled, Completed). Below the search bar is a 'Results' section with columns for Id, Created, Name, Agency, Owner, Status, Seeds, and Action. A green arrow points to the 'create new' button in the top right corner of the results section.

3. In the General tab, enter the following detail

Name: IIPC

Description: IIPC Website

State: Approved



The screenshot shows the 'General' tab of the Target setup form. The form includes fields for Id, Name (set to 'IIPC'), Description (set to 'IIPC Website'), Reference Number, Run on Approval, Use Automated QA, Owner (set to 'Workshop User'), State (set to 'Approved'), Auto-prune, Reference Crawl, and Request to Archivists. A green arrow points to the 'Name' field, another to the 'Description' text area, and a third to the 'State' dropdown menu.

4. Navigate to the Seeds tab, and enter the following detail, then click the Link button.

Seed: http://netpreserve.org/

Authorisation: Add Later

The screenshot shows the 'Targets' interface for 'IIPC'. The 'Seeds' tab is selected. In the 'General' section, the 'Seed' field contains 'http://netpreserve.org/' and the 'Authorisation' dropdown is set to 'Add Later'. A red box highlights the 'link' button. Two green arrows point from the 'Authorisation' dropdown and the 'link' button towards the 'link' button. Below the general section is a table with columns: Seed, Primary, Harvest Auth, Auth Agent, Start, End, Status, and Action. The 'Harvest Auth' column for the first row has a yellow warning icon. At the bottom are 'save' and 'cancel' buttons.

5. 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.

The screenshot shows the 'Targets' interface for 'IIPC'. The 'Seeds' tab is selected. In the 'General' section, the 'Seed' field contains 'http://netpreserve.org/**' and the 'Authorisation' dropdown is set to 'Auto'. A red box highlights the 'add' button in the 'Harvest Auth' column of the table. A green arrow points from the 'add' button towards the 'add' button. Below the table are 'save' and 'cancel' buttons.

6. 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.

The screenshot shows the 'Targets' interface for 'IIPC'. The 'Seeds' tab is selected. In the 'General' section, the 'Seed' field contains 'http://netpreserve.org/'. Below is a table titled 'Harvest Authorisations'. The 'Harvest Authorisation' column has two entries: 'netpreserve.org' (with a checked checkbox) and 'Blanket Authorisation' (with an unchecked checkbox). The 'Authorising Agent' column shows 'IIPC' for the first entry and 'Dummy Authorising Agency' for the second. The 'URL Patterns' column lists URLs for both. The 'Start' and 'End' columns show dates. The 'Status' column shows 'Approved' for both entries, with a yellow warning icon next to the first entry. At the bottom are 'done' and 'cancel' buttons.

7. 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*

The screenshot shows the 'Targets' interface with the 'Profile' tab active. The 'Base Profile' section is highlighted with a green border. Inside this section, the 'Harvester Type' dropdown is set to 'HERITRIX3' and the 'Base Profile' dropdown is set to 'Default - workshop'.

8. Navigate to the Schedule tab, click Create New.

The screenshot shows the 'Targets' interface with the 'Schedule' tab active. A green arrow points to the 'create new' button at the bottom right of the schedule table.

9. 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>

The screenshot shows the 'Targets' interface with a configuration dialog for scheduling. The dialog includes fields for 'From Date' (05/06/2019), 'To Date' (05/06/2019), 'Type' (Monthly), 'Time' (16:15), and 'Day of Month' (5). A green box highlights these input fields.

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

The screenshot shows the 'Targets' page with the 'Schedule' tab selected. There are two checkboxes: 'Harvest Now:' and 'Allow harvest optimization:'. Below is a table with one row:

Schedule	Owner	Next Scheduled Time	Action
15 16 5 * ? *	W. User	05/06/2019 16:15:00	

At the bottom are 'save' and 'cancel' buttons, with a large green arrow pointing to the 'save' button.

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

The screenshot shows the 'Target Instances' tab selected in the navigation bar. The main area is titled 'Target Summary' and contains a search form. The search form includes fields for ID, From, To, Agency, Owner, Name, Flag, State, and QA Recommendation. Buttons for 'search' and 'reset' are also present.

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.

The screenshot shows the 'Target Summary' page with the 'Results' table. The table has columns: Thumbnail, Name, Harvest Date, State, Owner, Run Time, Data Downloaded, URLs, % Failed, Crawls, QA Recom, and Action. One row is selected, and a green arrow points to the 'archive' button in the 'Action' column.

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 Edit Applications 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
Average URL/s:	0.04
Current KB/s:	0
Current URL/s:	0.11
URLs Downloaded:	2
URLs Queued:	2
URLs Failed:	0
Data Downloaded:	420 bytes
Elapsed Time:	00:00:00.47

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

The screenshot shows the 'Target Summary' interface for target ID 524288. The top navigation bar includes tabs for General, Profile, Harvest State, Logs (which is the active tab), and Display. Below the tabs is a table listing log files with columns for Filename, Action, and Size. The 'Action' column contains links for 'View' and 'Download'. A large green arrow points from the 'Logs' tab to the 'View' link for the 'crawlog' entry.

Filename	Action	Size
crawlog	View Download	0 bytes
nonfatal-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.

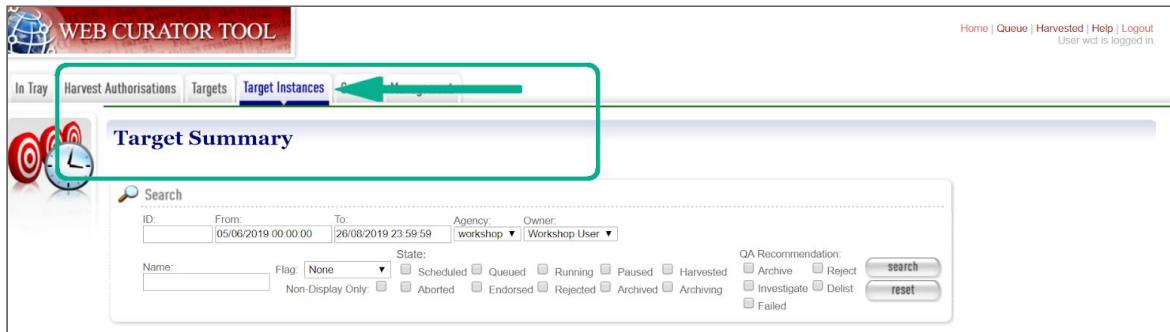
Log viewer: crawl.log		
2019-05-27T02:26:23.427z	1	55 dns:netpreserve.org F http://netpreserve.org/ text/dns #001 20190527022617467+4565 shal:1BMGZHNK65B7JYF5F6d5552ZLTCPSF^
2019-05-27T02:27:05.561z	200	67 http://netpreserve.org/robots.txt F http://netpreserve.org/ text/dns #001 20190527022705561+5370 shal:1S20B6L1K7ER0741
2019-05-27T02:27:29.382z	1	74 dns:www.googletagmanager.com E http://www.googletagmanager.com/etag/jstid= text/dns #009 20190527022729382+14 shal:1+4h
2019-05-27T02:27:29.405z	1	46 dns:fonts.googleapis.com E http://fonts.googleapis.com/ text/dns #002 20190527022729405+14 shal:3VWVZDKGULFCNKJSDR04H061AR-JL4LR0X
2019-05-27T02:27:29.404z	1	66 dns:fonts.googleapis.com EP http://fonts.googleapis.com/ text/dns #003 20190527022729404+14 shal:AF5W4R25M5BNPfA4Y43SW
2019-05-27T02:27:29.487z	1	50 dns:theme.co E http://theme.co/media/by-integrity-91.jpg text/dns #007 20190527022729487+14 shal:NC4U7HRC350KGDY2D351
2019-05-27T02:27:30.098z	200	103263 http://netpreserve.org/ - - text/dns #001 20190527022730098+14 shal:GQ304XSYOUTQXESD7E0XWGNHSLEIWPY - 3t
2019-05-27T02:27:32.581z	200	237 http://s.w.org/robots.txt EP http://s.w.org/ text/plain #009 20190527022732381+102 shal:Y4TNS7C5JSDIIY1XT2D3ENWHGEW!
2019-05-27T02:27:32.647z	200	25 http://fonts.googleapis.com/robots.txt EP http://fonts.googleapis.com/ text/plain #009 20190527022732647+102 shal:FB1B0J
2019-05-27T02:27:33.328z	404	157 https://www.googletagmanager.com/robots.txt EP https://www.googletagmanager.com/etag/jstid= text/html #003 20190527022733328+14
2019-05-27T02:27:35.693z	301	0 http://s.w.org/ E http://netpreserve.org/ text/html #009 20190527022735693+14 shal:31H2383NCFY2H7V7XZD7XQ8P - 3t
2019-05-27T02:27:35.953z	301	184 http://theme.co/robots.txt EP http://theme.co/media/by-integrity-91.jpg text/dns #017 20190527022735953+14 shal:1+4ZC21WH
2019-05-27T02:27:36.001z	1603	185 dns:www.googletagmanager.com E http://www.googletagmanager.com/etag/jstid= text/html #001 20190527022736001+14 shal:YHNU5eCY2H2X4J3O
2019-05-27T02:27:38.857z	200	65720 https://www.googletagmanager.com/etag/jstid= E http://netpreserve.org/ application/javascript #001 20190527022738857+14
2019-05-27T02:27:38.982z	200	4286 http://s.w.org/favicon.ico EI http://s.w.org/ image/x-icon #017 20190527022738982+104 shal:H4UGKGMNTVHQ3DKVLWEXFNHBBN
2019-05-27T02:27:39.003z	1	55 dns:www.gstatic.com EXP http://www.gstatic.com/www/loader.js text/dns #017 20190527022739003+104 shal:LBG6QYEB01V7UHIF
2019-05-27T02:27:39.027z	1	66 dns:adservice.google.com EXP https://adservice.google.com/dm/regclk/text/dns #017 20190527022739027+14 shal:MK3K5J3E3C
2019-05-27T02:27:39.045z	1	63 dns:ad.doubleclick.net EXP http://ad.doubleclick.net/activity.cnt/text/dns #017 20190527022739045+14 shal:RDLXPF5RNBDPIFPCJ
2019-05-27T02:27:39.069z	1	66 dns:www.googletraveladservices.com EXP https://www.googletraveladservices.com/travel/flights/cnt/text/dns #017 20190527022739069+14
2019-05-27T02:27:39.088z	1	64 dns:www.googleadservices.com EXP http://www.googleadservices.com/pageconversion_js/text/dns #017 20190527022739088+14 shal:1+4ZC21WH
2019-05-27T02:27:39.116z	74	74 dns:www.google-analytics.com EXP http://www.google-analytics.com/etag/jstid= text/html #001 20190527022739116+14 shal:1+4ZC21WH
2019-05-27T02:27:39.125z	200	139 dns:www.googletagmanager.com E http://www.googletagmanager.com/etag/jstid= text/html #001 20190527022739125+14 shal:1+4ZC21WH
2019-05-27T02:27:39.424z	1	66 dns:fonts.gstatic.com EEP http://fonts.gstatic.com/fonts/montserrat/v1/JTUStIgl_it63kChKm45W9lhrg.ttf text/dns #002 20190527022739424+14
2019-05-27T02:27:41.373z	404	1572 https://www.googletagmanager.com/favicon.ico EI https://www.googletagmanager.com/etag/jstid= text/html #017 20190527022741373+14
2019-05-27T02:27:41.406z	200	2238 https://www.gstatic.com/robots.txt F http://www.gstatic.com/robots.txt/text/html #001 20190527022741406+14 shal:1+4ZC21WH

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.

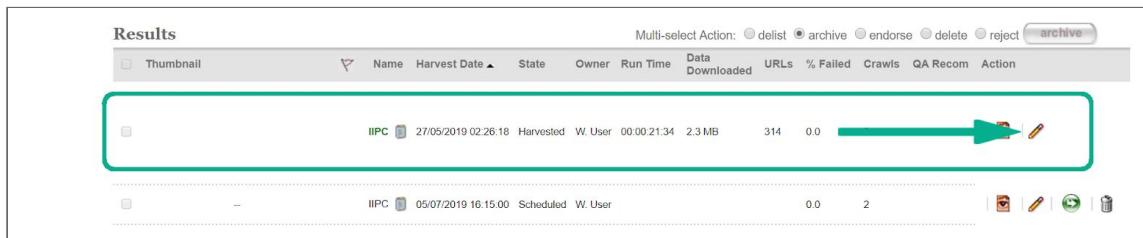
For more information on WCT quality review, see the [documentation](#).

1. Navigate to the Target Instance tab



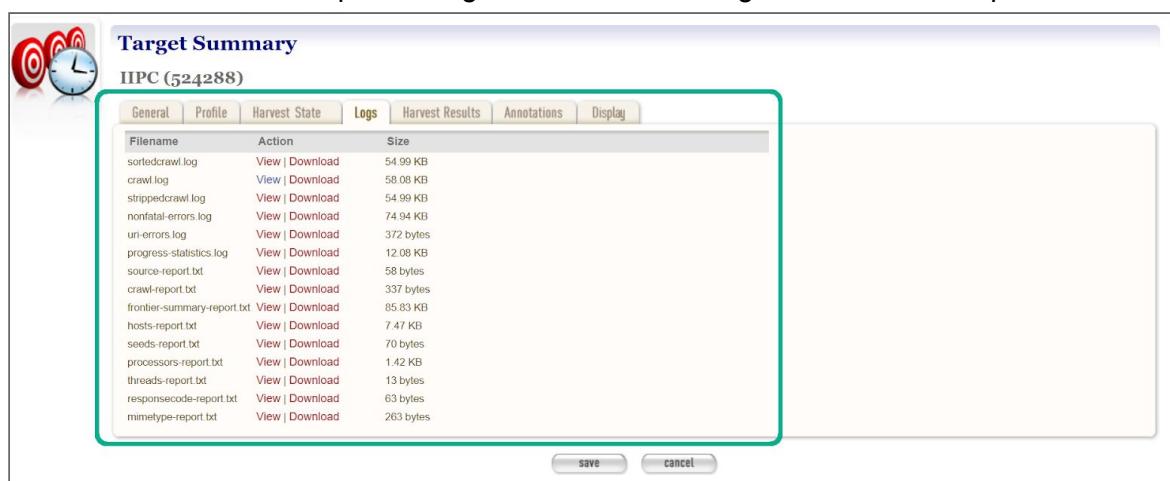
The screenshot shows the 'WEB CURATOR TOOL' interface. The top navigation bar includes links for Home, Queue, Harvested, Help, and Logout. The user 'wct' is logged in. Below the navigation is a toolbar with tabs: In Tray, Harvest Authorisations, Targets, and Target Instances (which is highlighted with a green arrow). A sidebar on the left features icons for targets and a clock. The main content area is titled 'Target Summary' and contains a search bar and a form for filtering target instances by ID, name, state, and QA recommendation. Buttons for 'search' and 'reset' are present.

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



The screenshot shows a 'Results' table with columns for Thumbnail, Name, Harvest Date, State, Owner, Run Time, Data Downloaded, URLs, % Failed, Crawls, QA Recom, and Action. One row is highlighted with a green arrow, representing a target instance named 'IIPC' that has been harvested. An edit icon (pencil) is visible in the Action column for this row.

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 screenshot shows the 'Target Summary' page for target instance 'IIPC (524288)'. The 'Logs' tab is selected, highlighted with a green arrow. A table lists various log files with their actions (View or Download) and sizes. The table includes rows for sortedcrawl.log, crawl.log, strippedcrawl.log, nonfatal-errors.log, uri-errors.log, progress-statistics.log, source-report.txt, crawl-report.txt, frontier-summary-report.txt, hosts-report.txt, seeds-report.txt, processors-report.txt, threads-report.txt, responsecode-report.txt, and mimetype-report.txt. At the bottom are 'save' and 'cancel' buttons.

Filename	Action	Size
sortedcrawl.log	View Download	54.99 KB
crawl.log	View Download	58.08 KB
strippedcrawl.log	View Download	54.99 KB
nonfatal-errors.log	View Download	74.94 KB
uri-errors.log	View Download	372 bytes
progress-statistics.log	View Download	12.08 KB
source-report.txt	View Download	58 bytes
crawl-report.txt	View Download	337 bytes
frontier-summary-report.txt	View Download	85.83 KB
hosts-report.txt	View Download	7.47 KB
seeds-report.txt	View Download	70 bytes
processors-report.txt	View Download	1.42 KB
threads-report.txt	View Download	13 bytes
responsecode-report.txt	View Download	63 bytes
mimetype-report.txt	View Download	263 bytes

4. 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.

The screenshot shows a log viewer window titled "Log viewer: hosts-report.txt". The main area displays a list of URLs with their respective counts of URLs and bytes. The list includes entries like "dns: 0 22 114 6449 0 0 0", "fonts.googleapis.com 0 2 12 16571 0 0 0", and various WordPress-related URLs such as "www.googletraveladservices.com 0 0 3 5518 0 0 0" and "wp-content/themes/2018/04/2018-04-16-16-56-6159 0 0 0". Below the list are several control buttons: "Show Line Numbers" (checkbox), "Number of lines to display" (text input set to 700), "Filter Type" (dropdown menu set to "Return the specified number of lines from the TAIL of the file"), an "apply" button, and a "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 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 0 6 4433 0 0 0
5 6159 s.w.org 16 5 6159 0 0 0
3 42794 wordpress.org 0 1 3 42794 0 0 0
3 5518 www.googletraveladservices.com 0 0 3 5518 0 0 0
2 2018 0 0 1 2018 0 0 0
2 4563 wf.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 arq.wordpress.org 0 0 2 33957 0 0 0
2 33899 aet.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 bce.wordpress.org 0 0 2 37446 0 0 0
2 33764 bci.wordpress.org 0 0 2 33764 0 0 0
2 33773 bre.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 700
Filter Type Return the specified number of lines from the TAIL of the file 


```

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

The screenshot shows a log viewer window titled "Log viewer: mimetype-report.txt". The main area displays a list of mime types with their counts. The list includes "text/plain" (36494), "text/html" (2034365), "text/css" (10054), "application/javascript" (198720), "image/x-icon" (10491), "application/rss+xml" (2590), "image/jpeg" (16783), "font/ttf" (46780), "image/gif" (290), and "text/javascript" (44915). Below the list are several control buttons: "Displaying: 100% of 263 B" and a "done" button.

```

Log viewer: mimetype-report.txt

[#files] [#bytes] [mime-types]
114 dns: 0 22 114 6449 0 0 0
102 36494 text/plain
78 2034365 text/html
7 10054 text/css
3 198720 application/javascript
3 10491 image/x-icon
2 2590 application/rss+xml
2 16783 image/jpeg
1 46780 font/ttf
1 290 image/gif
1 44915 text/javascript
Displaying: 100% of 263 B

done

```

6. 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.

The screenshot shows a "Target Summary" interface for target ID "IIPC (524288)". The "Harvest Results" tab is selected. A table below shows one harvest entry with details: No. 1, Date 27/05/2019 02:48:04, Derived From "W. User", Notes "Original Harvest", State "Review", and Action "Review". At the bottom are "save" and "cancel" buttons. A green arrow points to the "Review" link in the table.

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

7. 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.

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

Target Summary

IIPC (524288)

Quality Review Tools

Browse <http://netpreserve.org/>

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. You can browse and review the harvest here.

← → ⌂ localhost:3080/owb/wayback/20190527022723/http://netpreserve.org/ ⌂

OpenWayback

1 captures 27 May 19 - 27 May 19

APR MAY JUN 27 2018 2019 2020

Help ?

IIPC Menu

- [HOME](#)
- [ABOUT IIPC](#)
 - [IIPC members](#)
 - [Leadership](#)
 - [Working groups](#)
 - [Projects](#)
 - [Resources](#)
 - [Join the IIPC](#)
 - [IIPC mailing list](#)
 - [Close](#)
- [WEB ARCHIVING](#)
 - [About archiving](#)
 - [Video spotlights](#)
 - [Tools & software](#)
 - [OpenWayback](#)
 - [Legal issues](#)
 - [Legal deposit](#)
 - [Collection development policies](#)
 - [Case studies](#)
 - [Bibliography](#)
 - [Close](#)

9. 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](#).

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

Target Summary

IIPC (524288)

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

No.	Date	Derived From	User	Notes	State	Action
1	27/05/2019 02:48:04		W. User	Original Harvest		Review Endorse Reject for Reason: <input type="button" value="▼"/>

save cancel

Tutorial B - Crawl Scoping

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

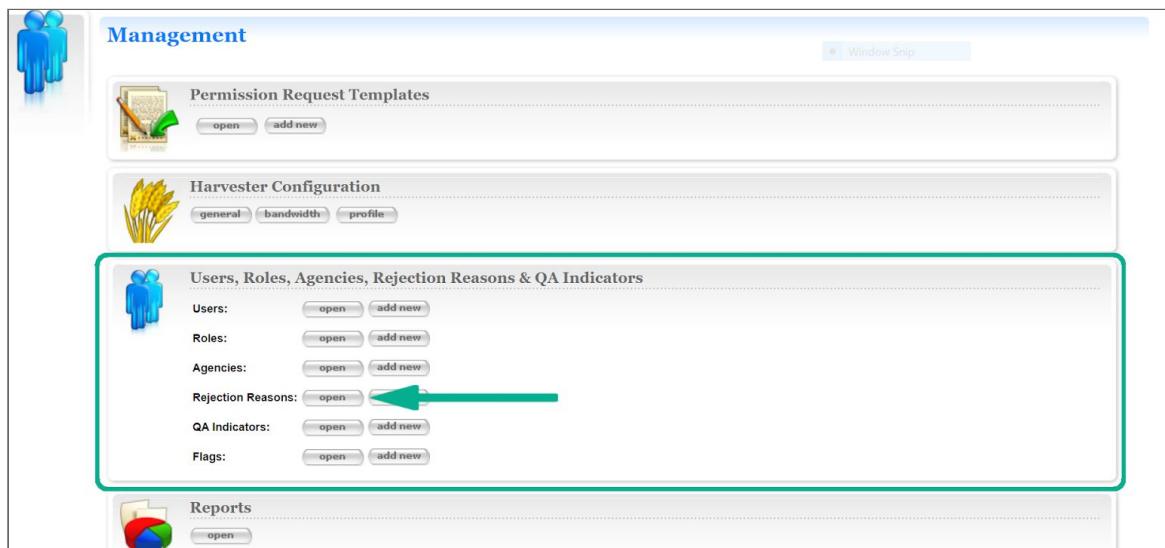
Scenario: You must crawl a website that is under a domain that your institution has authorisation to harvest. Eg. ccTLD like .uk, .eu, .nz or institution like harvard.edu

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>

A screenshot of the 'Users, Roles, Agencies & Rejection Reasons' creation form. The top navigation bar includes 'In Tray', 'Harvest Authorisations', 'Targets', 'Target Instances', 'Groups', and 'Management'. The 'Management' tab is active. The form has a title 'Users, Roles, Agencies & Rejection Reasons' and a sub-section 'Rejection Reason'. It contains fields for 'Agency' (set to 'workshop'), 'Rejection Reason' (input field containing 'No new content'), and two checkboxes: 'Available as a Rejection Reason for Targets' (unchecked) and 'Available as a Rejection Reason for harvested Target Instances' (checked). At the bottom are 'save' and 'cancel' buttons.

- 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>

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](#).

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

Id:	
Name:	NLNZ *
Description:	National Library of New Zealand website
Reference Number:	
Run on Approval:	<input type="checkbox"/>
Use Automated QA:	<input type="checkbox"/>
Owner:	Workshop User
State:	Approved
Auto-prune:	<input type="checkbox"/>
Reference Crawl:	<input type="checkbox"/>
Request to Archivists:	

4. 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

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="import"/>							

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

Harvester Type: HERITRIX3

Base Profile: Default - workshop

6. 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>

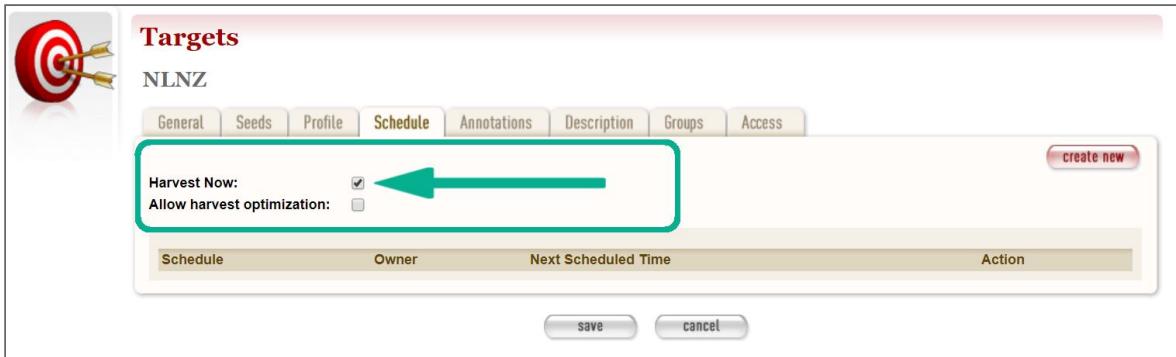
Subject: GLAM

Type: Collection

Language: <language of your institution website>

Title:	NLNZ
Identifier:	
Description:	National Library of New Zealand
Subject:	GLAM
Creator:	
Publisher:	
Contributor:	
Type:	Collection
Format:	
Source:	
Language:	English
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.



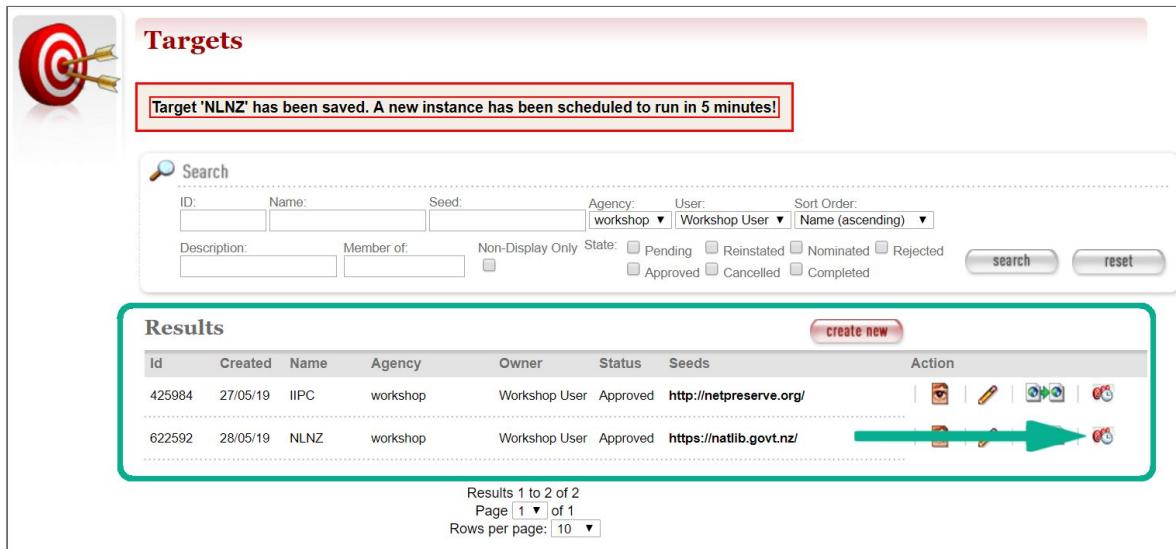
The screenshot shows the 'Targets' interface with the 'NLNZ' target selected. The 'Schedule' tab is active. A green box highlights the 'Harvest Now' checkbox, which is checked, and a green arrow points to it from the left. Below it is another checkbox for 'Allow harvest optimization'. At the bottom are 'save' and 'cancel' buttons.

- 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.



The screenshot shows the 'Targets' interface with the 'NLNZ' target selected. The 'Results' section is highlighted with a green box. It lists two target instances: one for 'IIPC' and one for 'NLNZ'. For the 'NLNZ' instance, a green arrow points to the 'Harvest Now' icon in the 'Action' column. At the bottom, there are pagination controls and a 'Rows per page' dropdown set to 10.

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



The screenshot shows the 'Results' interface with a single target instance listed. The 'Action' column for this instance contains a green arrow pointing to the 'Harvest Now' icon. At the bottom, there are pagination controls and a 'Rows per page' dropdown set to 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

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: <input type="text"/>	From: <input type="text"/> 28/05/2019 10:00:00	To: <input type="text"/>	Agency: <input type="button"/>	Owner: <input type="button"/>
Name: <input type="text"/> NLNZ	Flag: <input type="button"/> None	<input type="checkbox"/> Scheduled <input type="checkbox"/> Queued <input type="checkbox"/> Running <input type="checkbox"/> Paused <input type="checkbox"/> Harvested	<input type="checkbox"/> Archive <input type="checkbox"/> Reject	search
<input type="checkbox"/> Non-Display Only <input type="checkbox"/> Aborted <input type="checkbox"/> Endorsed <input type="checkbox"/> Rejected <input type="checkbox"/> Archived <input type="checkbox"/> Archiving <input type="checkbox"/> Failed				

Results

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	0										

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

6. Navigate to the Logs tab, and click on the View link for crawl.log.
7. 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

```

2019-05-28T10:07:55.4452 301 194 https://embed.ly/code?url=LREX https://cdn.embedly.com/widgets/platform.js text/html #019
2019-05-28T10:07:55.4462 200 5093 https://twitter.com/intent/tweet LREX https://cdn.embedly.com/widgets/platform.js text/html
2019-05-28T10:07:55.4642 302 0 https://www.pinterest.nz/pin/create/button/ LREX https://cdn.embedly.com/widgets/platform.js text/html #007
2019-05-28T10:07:56.0102 301 76 dns:www.pinterest.nz LREXRP https://www.pinterest.nz/pin/create/button/ text/dns #014 2019
2019-05-28T10:07:56.4872 200 185 https://api-cdn.embed.ly/ LREX https://cdn.embedly.com/widgets/platform.js text/html #007
2019-05-28T10:07:56.4872 200 4632 https://natlib.govt.nz/kakapo/assets/favicon-ca1352313f0ac0a33b4934eb70a5f05c4376f93c6a10a!
2019-05-28T10:07:58.3622 200 25726 https://images.ctfassets.net/n3pl04aq7k6m/2z5leE2zm4ACEMcc6o1O/3ac0c4465fad4d0e4e95e00a8?
2019-05-28T10:07:58.4532 200 706 https://www.youtube.com/robots.txt LREXRF https://www.youtube.com/watch?v= text/plain #015
2019-05-28T10:07:58.4632 200 2856 https://app.embed.ly/ LREX https://cdn.embedly.com/widgets/platform.js text/html #005 2019
2019-05-28T10:07:58.5092 200 1037 https://vimeo.com/robots.txt LREXRP https://vimeo.com/ text/plain #020 2019052810075811243
2019-05-28T10:07:58.5662 200 184 https://cdn.embedly.com/widgets/embed?url= LREX https://cdn.embedly.com/widgets/platform.js
2019-05-28T10:07:58.5662 200 9377 https://www.pinterest.nz/robots.txt LREXRP https://www.pinterest.nz/pin/create/button/ tex
2019-05-28T10:07:59.4662 404 178 https://embed.ly/robots.txt LREXRP https://embed.ly/code?url= text/html #019 201905281007
2019-05-28T10:07:59.7022 200 4632 https://natlib.govt.nz/kakapo/assets/favicon-ca1352313f0ac0a33b4934eb70a5f05c4376f93c6a10a!
2019-05-28T10:08:01.4552 200 21256 https://soundcloud.com/track/ LREX https://cdn.embedly.com/widgets/platform.js text/html #
2019-05-28T10:08:01.7922 404 60 https://images.ctfassets.net/favicon.ico LREI https://images.ctfassets.net/n3pl04aq7k6m/1F?
2019-05-28T10:08:01.8492 301 0 https://www.youtube.com/watch?v= LREX http://www.youtube.com/watch?v= text/html #015 2019
2019-05-28T10:08:02.0682 200 25546 https://vimeo.com/ LREXRP http://vimeo.com/ text/html #020 20190528100801514+52 sha1:272821
2019-05-28T10:08:02.1302 403 243 https://cdn.embedly.com/widgets/platform.js applicat
2019-05-28T10:08:02.9512 301 0 https://www.pinterest.nz/pin/create/button/ LREX https://www.pinterest.com/pin/create/but
2019-05-28T10:08:05.5372 1 242 https://natlib.govt.nz/robots.txt LXP http://natlib.govt.nz/kakapo/assets/wellsprint_logo-c
2019-05-28T10:08:05.5392 200 13825 https://embed.ly/code?url= LREX http://embed.ly/code?url= text/html #025 2019052810080452?
2019-05-28T10:08:06.0862 200 4632 https://natlib.govt.nz/kakapo/assets/wellsprint_logo-ca1352313f0ac0a33b4934eb70a5f05c4376f9
2019-05-28T10:08:07.7672 200 546371 https://www.youtube.com/ LREXRR https://www.youtube.com/watch?v= text/html #015 2019052810

```

Show Line Numbers

Number of lines to display

Filter Type

8. 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

```

2019-05-28T10:07:48.1062 1 76 dns:www.pinterest.com LREXP https://www.pinterest.com/pin/create/button/ text/dns #016 20190
2019-05-28T10:07:48.2002 1 76 dns:log.pinterest.com LREXP https://log.pinterest.com/ text/dns #016 20190528100748159+41 sha
2019-05-28T10:07:51.9042 200 0 https://log.pinterest.com/robots.txt LREXP https://log.pinterest.com/ unknown #003 2019052810
2019-05-28T10:07:52.0262 302 0 https://www.pinterest.com/robots.txt LREXP https://www.pinterest.com/pin/create/button/ text/
2019-05-28T10:07:55.3052 200 0 https://log.pinterest.com/ LREX https://cdn.embedly.com/widgets/platform.js unknown #003 2019
2019-05-28T10:07:55.4642 302 0 https://www.pinterest.com/pin/create/button/ LREX https://cdn.embedly.com/widgets/platform.js
2019-05-28T10:07:55.5372 1 9377 https://www.pinterest.nz LREXRP https://www.pinterest.nz/pin/create/button/ text/dns #014 201905
2019-05-28T10:07:59.0722 200 76 dns:www.pinterest.nz LREXRP https://www.pinterest.nz/pin/create/button/ text/dns #014 201905
2019-05-28T10:08:02.5802 302 0 https://www.pinterest.nz/pin/create/button/ LREXR https://www.pinterest.com/pin/create/button/
2019-05-28T10:08:05.5832 -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

9. 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](#).

10. Navigate back to the Target Instances tab, tick the Running state filter, and click Search.

The screenshot shows the 'Target Summary' page. At the top, there's a search bar with fields for ID, From (28/05/2019 10:00:00), To, Agency, and Owner. Below the search bar is a 'State' filter section with checkboxes for Scheduled, Queued, Running, Paused, Harvested, Aborted, Endorsed, Rejected, Archived, Archiving, and Failed. The 'Running' checkbox is checked. To the right of the filter are 'QA Recommendation' checkboxes for Archive, Reject, Investigate, Delist, and Failed. Below the filter is a 'Results' table with columns: Thumbnail, Name, Harvest Date, State, Owner, Run Time, Data Downloaded, URLs, % Failed, Crawls, QA Recom, and Action. One row is shown: NLNZ, 28/05/2019 10:05:50, Running, W User, 00:00:30:34, 46.04 MB, 889, 0.0, 3. The action column contains icons for edit, archive, endorse, delete, and reject. A green arrow points to the 'Running' checkbox in the filter, and another green arrow points to the 'Search' button.

11. 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 the same search and filter settings as the previous screenshot. The results table shows one row for NLNZ. In the 'Action' column for this row, there is a red circle with a white X icon, which is the stop icon. A green arrow points to this icon.

12. 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.

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

The screenshot shows the 'Targets' tab with a 'Results' table. The table has columns: Id, Created, Name, Agency, Owner, Status, Seeds, and Action. Two rows are shown: one for IIPC with ID 425084 and one for NLNZ with ID 622592. The NLNZ row is highlighted with a green border. The 'Action' column for the NLNZ row contains icons for edit, archive, endorse, delete, and reject. A green arrow points to the edit icon (pencil) in the 'Action' column for the NLNZ target.

3. 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.**

Profile Element	Override Value	Enable Override
Document Limit		<input type="checkbox"/>
Data Limit	0.0 <input type="button" value="B"/>	<input type="checkbox"/>
Time Limit	0.0 <input type="button" value="SECOND"/>	<input type="checkbox"/>
Max Path Depth		<input type="checkbox"/>
Max Hops	4 <input type="text"/>	<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 type="text"/>	<input checked="" type="checkbox"/>
Include URLs		<input type="checkbox"/>

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

Compare and review the crawl

1. 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

Filter Type

Regular Expression

2. 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.

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

The screenshot shows a search interface for target instances. At the top, there are fields for ID (1048576), From, To, Agency, and Owner. Below these are dropdowns for State (e.g., Scheduled, Queued, Running, Paused, Harvested) and QA Recommendation (Archive, Reject). A search button and a reset button are also present. The main area is titled "Results" and contains a table with columns: Thumbnail, Name, Harvest Date, State, Owner, Run Time, Data Downloaded, URLs, % Failed, Crawls, QA Recom, and Action. One row is highlighted with a green arrow pointing to it, representing the instance "NLNZ". At the bottom, there are links for "archive", "endorse", "delete", and "reject". Below the table, it says "Results 1 to 1 of 1", "Page 1 of 1", and "Rows per page: 10".

4. 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](#).

The screenshot shows the "Target Summary" screen for instance "NLNZ (1048576)". It includes sections for "Harvest Results", "Profile Overrides", "Resources", "Key Indicators" (which is highlighted with a green box), "Recommendation", "Schedule", "Annotations", and "Harvest History". The "Harvest History" section at the bottom shows a table of previous harvest runs:

Start Date	Status	Data	URLs	Failed	Elapsed	KB/s	Job Status	QA Phase
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

At the bottom right, there is a "Add Annotation" section with a text input field and a "add annotation" button.

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

The screenshot shows a table titled "Results" with one row highlighted. The columns include: Thumbnail, Name (NLNZ), Harvest Date (31/05/2019 09:39:19), State (Harvested), Owner (W. User), Run Time (00:00:09:49), Data Downloaded (12.32 MB), URLs (273), % Failed (0.0), Crawls, QA Recom, and Action. A green arrow points to the edit icon (pencil) in the Action column.

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

The screenshot shows the "Target Summary" page for target "NLNZ (1048576)". The "Harvest Results" tab is active. A table lists one harvested target with columns: No., Date, Derived From, User, Notes, State, and Action. The "Action" column contains links: Review, Endorse, and Reject for Reason. A green arrow points to the "Review" link. Below the table are "save" and "cancel" buttons.

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

The screenshot shows the "Quality Review Tools" page. It lists two tools: "Harvest History" and "Tree View". "Harvest History" is selected, indicated by a green arrow. The "Description" for Harvest History is: "Compare current harvest result with previous harvests." The "Description" for Tree View is: "Graphical view of harvested data." Below the tools is a "done" button.

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.

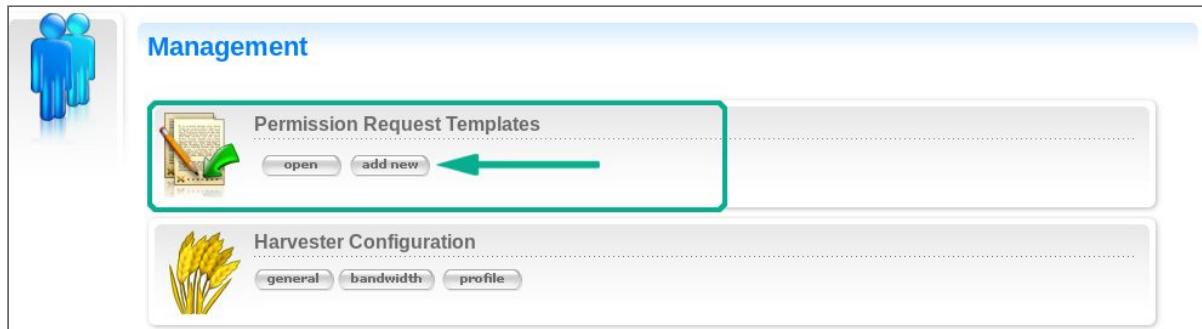
Create a Permission Request Template

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: *

Template Description:

Agency: ←

Template type:

Template Subject: *

Template Overwrite From:

Template From Address:

Template CC Address(s):

Template BCC Address(s):

Template Reply-to Address:

Template 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

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.

The screenshot shows the 'WEB CURATOR TOOL' interface with the 'Harvest Authorisations' tab selected. The search bar includes fields for ID, Name, Authorising Agent, Order No., Agency, Sort Order, and Permissions Status. The results table lists one item: 'Blanket Authorisation' created on '07/05/19' by 'Dummy Authorising Agency' with status 'Approved'. A green arrow highlights the 'create new' button in the search bar, which corresponds to the button in the results table.

3. Under the General tab enter:

Title: Advanced Authorisation

The screenshot shows the 'General' tab of the 'Harvest Authorisations' edit form. The 'Title' field is highlighted with a red box and contains the value 'Advanced Authorisation', which has a red asterisk (*) next to it, indicating it is a required field. Other fields visible include 'Agency' (workshop), 'Description', 'Order No.', 'Published' (unchecked), and 'Enabled' (checked). Below the form is an 'Annotations' section with an 'add' button, a table showing no annotations, and buttons for 'save' and 'cancel' at the bottom.

4. 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/>.

The screenshot shows the 'Harvest Authorisations' interface with the 'Advanced Authorisation' tab selected. The 'URL Patterns' tab is highlighted. A green box highlights the 'New URL Pattern:' input field containing 'https://www.kb.nl/*' and the 'add' button to its right. Below this is a table with columns 'URL Pattern' and 'Action'. At the bottom are 'save' and 'cancel' buttons.

5. Create a new Authorising Agency under the next tab.

The screenshot shows the 'Harvest Authorisations' interface with the 'Advanced Authorisation' tab selected. The 'Authorising Agencies' tab is highlighted. A green box highlights the 'create new' button. Below it is a table with columns 'Authorising Agency', 'Contact', and 'Action'. At the bottom are 'save' and 'cancel' buttons.

6. 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.

The screenshot shows a form for creating a new Authorising Agency. The 'Name:' field contains 'Koninklijke Bibliotheek' with a red asterisk indicating it's required. The 'Contact:' field contains 'Jane Doe' with a red asterisk. The 'Email:' field contains 'jane.doe@test.email.address'. Below the form are 'save' and 'cancel' buttons.

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

Harvest Authorisations

Advanced Authorisation

General URL Patterns Authorising Agencies Permissions

Status Date Requested Authorising Agent From To URL Patterns Action

create new

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>

Authorising Agent:	Koninklijke Bibliotheek										
Dates:	28/05/2019 * dd/mm/yyyy										
Status:	Approved										
Auth. Agency Response:	(empty text area)										
Special Restrictions:	(empty text area)										
Copyright Statement:	(empty text area)										
Copyright URL:	(empty text area)										
Access Status:	Open (unrestricted) access										
Open Access Date:	(empty text area)										
Quick Pick:	<input checked="" type="checkbox"/>										
Display Name:	Koninklijke Bibliotheek										
Urls:	<input checked="" type="checkbox"/> https://www.kb.nl/*										
File Reference:	(empty text area)										
Assign Approval Task:	No										
Exclusions											
URL	Reason										
(empty text area)											
add											
No exclusions have been defined.											
Annotations											
(empty text area)											
add											
<table border="1"> <thead> <tr> <th>Date</th> <th>User</th> <th>Notes</th> <th>Action</th> </tr> </thead> <tbody> <tr> <td colspan="4">EN_US No Annotations Available.</td> </tr> </tbody> </table>				Date	User	Notes	Action	EN_US No Annotations Available.			
Date	User	Notes	Action								
EN_US No Annotations Available.											
save		cancel									

9. Finally, save the Harvest Authorisation.



The screenshot shows the 'Harvest Authorisations' interface. A green checkmark icon is visible in the top-left corner. The main title is 'Harvest Authorisations'. Below it, a sub-section titled 'Advanced Authorisation' is selected. A navigation bar at the top includes tabs for 'General', 'URL Patterns', 'Authorising Agencies', and 'Permissions', with 'Permissions' being the active tab. A red 'create new' button is located in the top right. The main content area displays a table with one row of data:

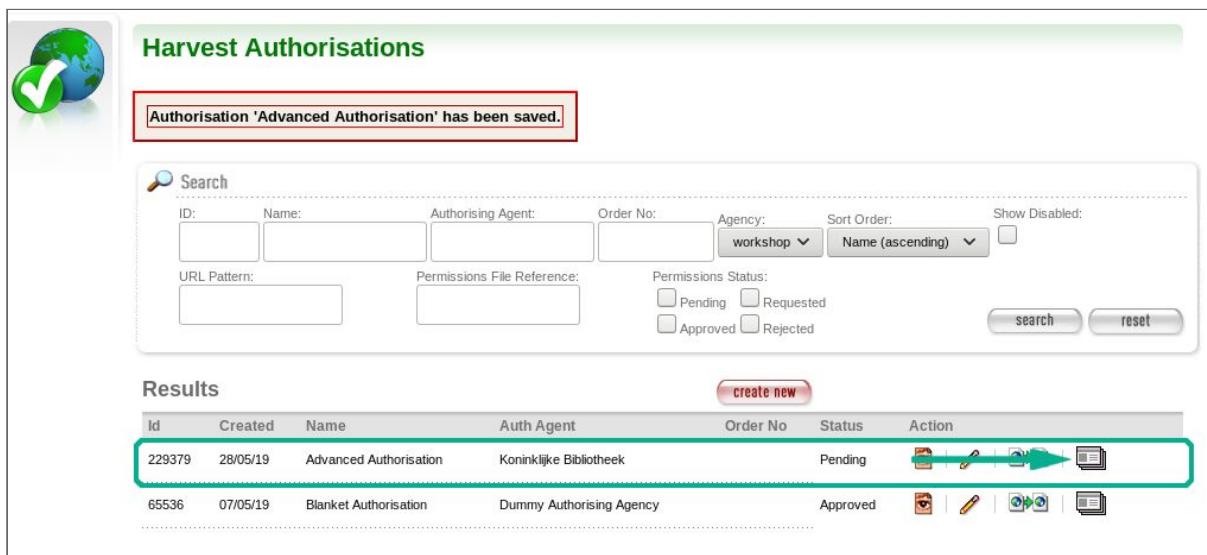
Status	Date Requested	Authorising Agent	From	To	URL Patterns	Action
Pending		Koninklijke Bibliotheek	28/05/2019		https://www.kb.nl/*	

Below the table is a control panel with a large green arrow pointing right, followed by 'save' and 'cancel' buttons.

Send Permission Request

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

- 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.



The screenshot shows the 'Harvest Authorisations' interface. A green checkmark icon is visible in the top-left corner. The main title is 'Harvest Authorisations'. A message box at the top center says 'Authorisation 'Advanced Authorisation' has been saved.' Below it is a search form with fields for ID, Name, Authorising Agent, Order No., Agency, Sort Order, and search buttons. The search results table shows two rows of data:

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	

- 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.

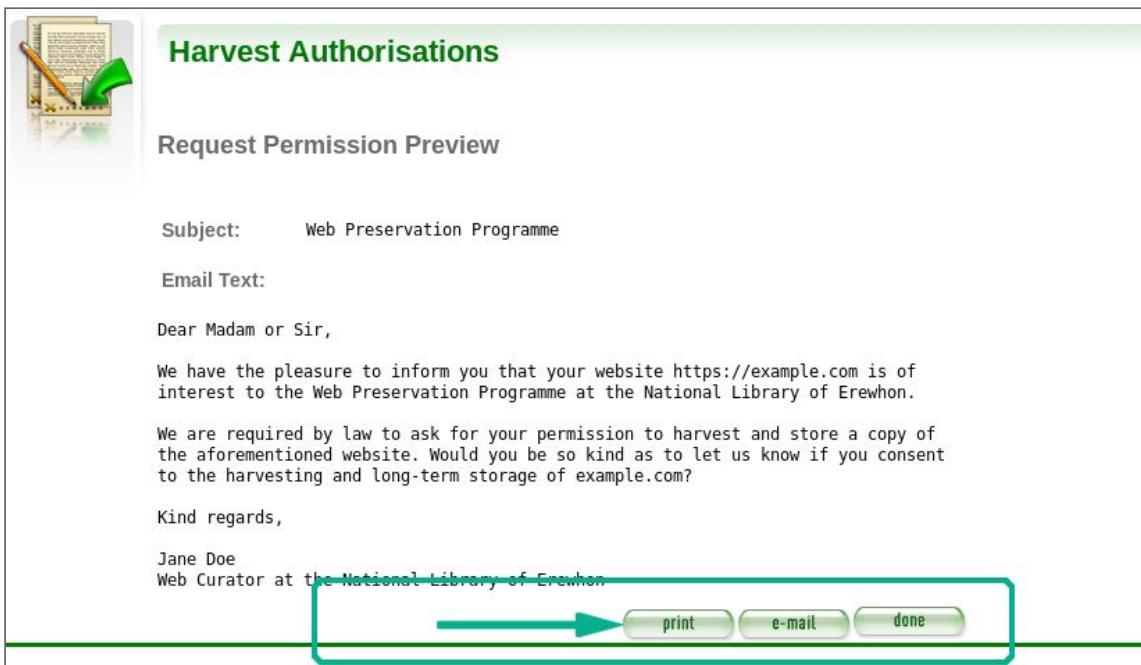


The screenshot shows the 'Harvest Authorisations' interface. A green checkmark icon is visible in the top-left corner. The main title is 'Harvest Authorisations'. The URL patterns table shows one row of data:

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

A red arrow points to the 'Template' dropdown menu, which is set to 'Generic Permission Request Template'.

- 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.



The screenshot shows a 'Request Permission Preview' window with the following content:

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.

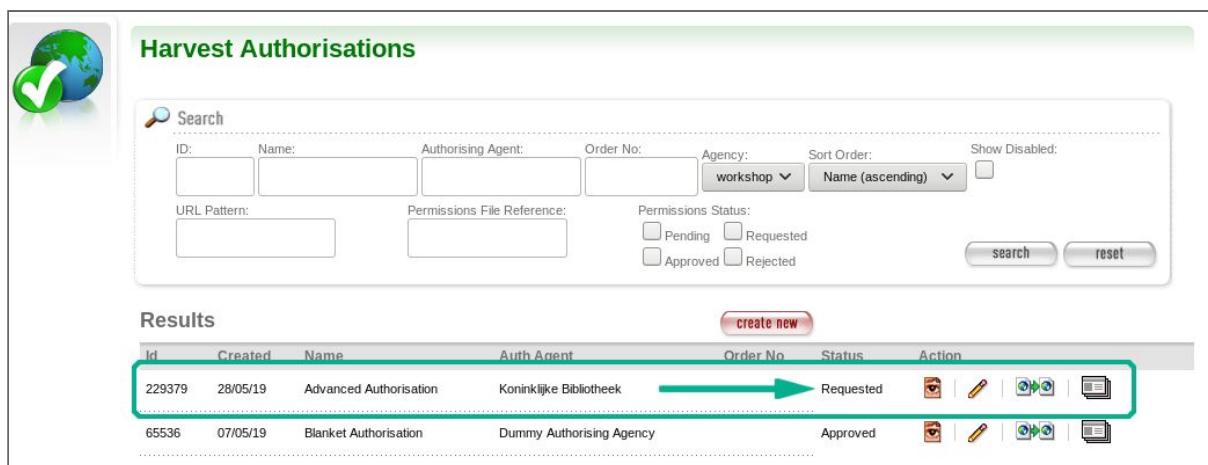
We are required by law to ask for your permission to harvest and store a copy of the aforementioned website. Would you be so kind as to let us know if you consent to the harvesting and long-term storage of example.com?

Kind regards,

Jane Doe
Web Curator at the National Library of Erewhon

At the bottom right, there are three buttons: **print**, **e-mail**, and **done**. A green arrow points from the text area towards the **print** button.

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



The screenshot shows a 'Harvest Authorisations' search results page with the following details:

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
- Buttons: **search** and **reset**

Results:

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						
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

General

Id: 589824
Agency: workshop

Title: Advanced Authorisation *

Description:

Order No:

Published:

Enabled:

Annotations

Date **User** **Notes** **Action**

There are no annotations available.

save **cancel**

4. Open the existing permission in edit mode.

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

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

Authorising Agent: Koninklijke Bibliotheek

Dates: 01/07/2019

Status: Approved

Auth. Agency Response:

Dear Ms. Doe,
Thank you for your email about the safeguarding of our institutional website. We are happy to grant you permission to harvest the site provided the archival copy will only be accessible on your premises.

Special Restrictions:

- * accessible on-site
- * no crawling before 1 July 2019

Copyright Statement:

Copyright URL:

Access Status: Restricted by location

Open Access Date:

Quick Pick:

Display Name: Koninklijke Bibliotheek

Urls: https://www.kb.nl/* *

File Reference:

Assign Approval Task: No

Exclusions

URL	Reason

No exclusions have been defined.

Annotations

Date	User	Notes	Action
EN_US No Annotations Available.			

6. Finally, save the Harvest Authorisation.

The screenshot shows the 'Harvest Authorisations' page with a green checkmark icon. The title 'Harvest Authorisations' is at the top. Below it is a section titled 'Advanced Authorisation' with tabs: General, URL Patterns, Authorising Agencies, and Permissions (which is selected). A 'create new' button is in the top right. A table lists one authorisation entry:

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

At the bottom are 'save' and 'cancel' buttons, with a large green arrow pointing 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' page with a red bullseye icon. The title 'Targets' is at the top. Below it is a table with tabs: General, Seeds, Profile, Schedule, Annotations, Description, Groups, and Access (General is selected). A 'KB' entry is listed in the table. The 'General' tab details for this target are shown in a box:

Name:	KB
Description:	Koninklijke Bibliotheek website
State:	Approved

Three green arrows point from the 'Name', 'Description', and 'State' fields to the corresponding fields in the 'General' tab of the Targets interface, indicating they are being populated from the Harvest Authorisation.

- 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.

Targets
KB

General Seeds Profile Schedule Annotations Description Groups Access

Seed: https://www.kb.nl/

Authorisation: Auto

link import

Seed	Primary	Harvest Auth	Auth Agent	Start	End	Status	Action
<input type="checkbox"/>							

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

Targets
KB

General Seeds Profile Schedule Annotations Description Groups Access

create new

Harvest Now:

Allow harvest optimization:

Schedule	Owner	Next Scheduled Time	Action
----------	-------	---------------------	--------

save cancel

- 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.

WEB CURATOR TOOL

In Tray Harvest Authorisations Targets Target Instances Groups Management

Targets

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

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

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

The screenshot shows the 'Target Summary' interface. At the top, there is a search bar with fields for ID, From, To, Agency, and Owner. Below the search bar are various filter options for Name, State (Flag: None, Scheduled, Queued, Running, Paused, Harvested), and QA Recommendation (Archive, Reject, Delist). A results table below shows one item: KB (786433) scheduled for harvest on 29/05/2019 at 14:26:06 by W. User. The table includes columns for Thumbnail, Name, Harvest Date, State, Owner, Run Time, Data Downloaded, URLs, % Failed, Crawls, QA Recom, and Action.

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

The screenshot shows the 'Harvest Now' dialog for target KB (786433). It displays the target details (Id: 786433, Target Name: KB, Schedule: 30/05/2019 14:26:06) and a table for assigning harvest agents. The table has columns for Harvest Agent, Max Harvests, and Current Harvests. An H3 Agent is listed with Max Harvests set to 2 and Current Harvests set to 0. A red arrow points to the 'allocate' button, which is highlighted with a green border. A 'done' button is located at the bottom right.

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.

The screenshot shows the 'Target Summary' interface again. A red box highlights a validation error message: 'The following validation errors have occurred. The target 720896 is not approved for harvest.' Below this, the 'Harvest Now' dialog is shown for target KB (786433). The error message is also present here. The harvest agent table shows the same information as before, with the 'allocate' button still highlighted by a green border.

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.

The screenshot shows the 'Harvester Configuration' interface with the 'Profile' tab selected. On the left, there's a sidebar with a wheat icon and a 'Harvester Configuration' title. Below it, there's a form for importing profiles from XML files, with fields for 'Select XML File' (a 'Choose File' button), 'Profile name' (a text input), 'Import to agency' (a dropdown set to 'workshop'), and 'Type to import' (a dropdown set to 'HERITRIX3'). A large 'import' button is at the bottom of this section. To the right, there's a toolbar with buttons for 'create new', 'refresh', and other actions. Below these sections is a table with columns: 'Name', 'Default', 'Description', 'Type', 'Status', 'Agency', and 'Action'. A single row is visible: 'Default - workshop' (Default), 'Default profile created by new agency action' (Description), 'HERITRIX3' (Agency), and 'Active' (Status). The 'Action' column for this row contains several icons, with a green arrow pointing to the first icon (likely the export icon).

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

The screenshot shows the WCT interface with the 'Management' tab selected. At the top, there's a navigation bar with links like 'In Tray', 'Harvest Authorisations', 'Targets', 'Groups', 'Target Instances', 'Reports', and 'Management'. Below this is a search bar with filters for 'Show Inactive Profiles', 'Agency Filter' (set to 'workshop'), and 'Type Filter'. A table lists profiles, with one row highlighted: 'Default - workshop' (Default), 'Default profile created by new agency action' (Description), 'HERITRIX3' (Agency), and 'Active' (Status). The 'Action' column for this row contains several icons. At the bottom, a message box shows a download progress bar for 'ExportedProfile.xml' (100% complete) with a green arrow pointing to the download icon. There are also 'Show all' and 'X' buttons in the message box.

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

```

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
ExportedProfile.xml
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
12 <context:annotation-config/>
13
14 <!--
15   OVERRIDES
16   Values elsewhere in the configuration may be replaced ('overridden')
17   by a Properties map declared in a PropertiesOverrideConfigurer,

```

Normal text file length : 193 lines : 3 Ln:2 Col:121 Sel:0|0 Windows (CR LF) UTF-8 INS

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.

- 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>
```

- 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" /> -->
```

- 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="" /> -->
```

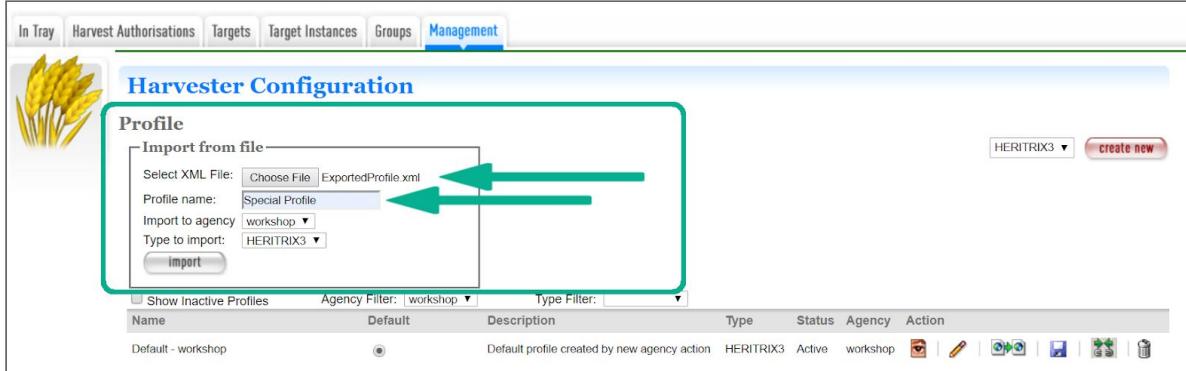
- Once you have edited and saved the profile xml, import it back into WCT, using the following details.

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

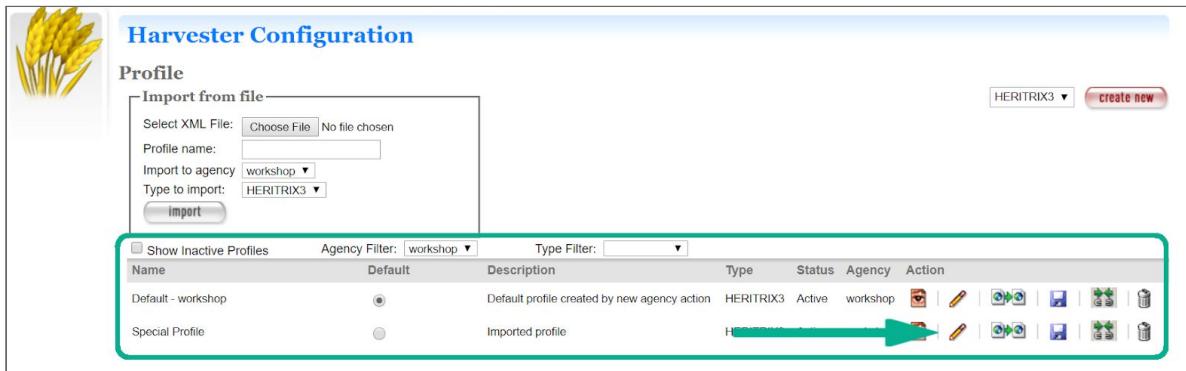
Profile name: Special profile

Import to agency: workshop

Type to import: HERITRIX3

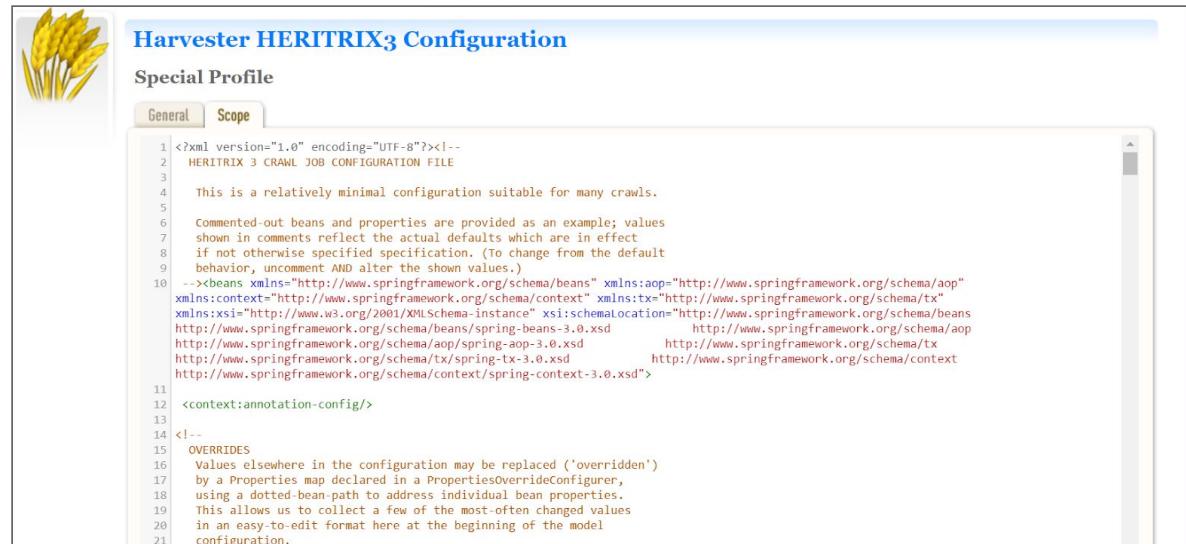


8. You should see the new imported profile listed - "Special Profile". Click the edit icon next to it.



9. 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.

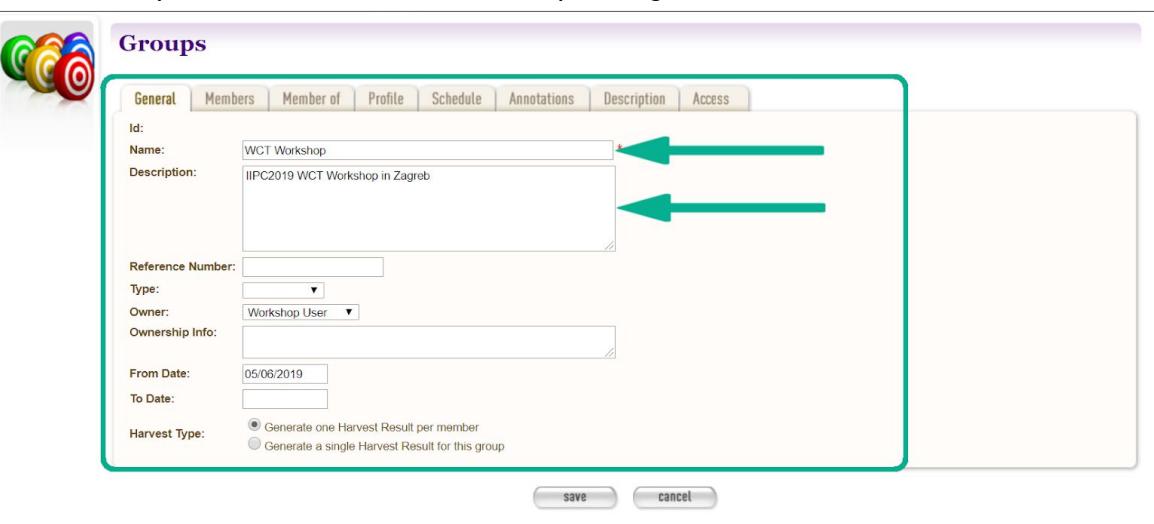
For more information on Groups in WCT, see the [documentation](#).

1. Navigate to the Groups tab within WCT.



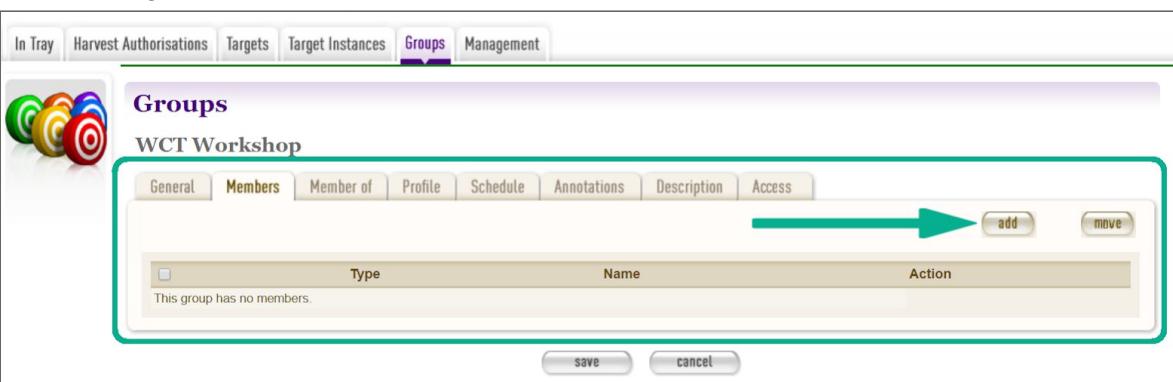
The screenshot shows the 'Groups' search interface. At the top, there's a navigation bar with tabs: In Tray, Harvest Authorisations, Targets, Target Instances, Groups (which is highlighted with a green arrow), and Management. Below the navigation bar is a search form titled 'Search'. It includes fields for ID, Name, Agency, Owner, Member of, Type, and Non-Display Only. There are also 'search' and 'reset' buttons. The main area is titled 'Groups' and features a decorative icon of three colorful targets.

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



The screenshot shows the 'General' tab of the group configuration. The 'Name' field contains 'WCT Workshop' and the 'Description' field contains 'IIPC2019 WCT Workshop in Zagreb'. Other fields visible include Reference Number, Type, Owner, Ownership Info, From Date (05/06/2019), To Date, and Harvest Type (radio buttons for 'Generate one Harvest Result per member' and 'Generate a single Harvest Result for this group'). At the bottom are 'save' and 'cancel' buttons.

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



The screenshot shows the 'Members' tab of the group configuration. The 'Members' tab is highlighted with a green arrow. At the top right of the tab area is an 'add' button, which is also highlighted with a green arrow. Below the tab is a table with columns: Type, Name, and Action. The table currently displays the message 'This group has no members.' At the bottom are 'save' and 'cancel' buttons.

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

Name	Type	Status	Agency
NLNZ	Target	Completed	workshop
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

General	Seeds	Profile	Schedule	Annotations	Description	Groups	Access
Id: Name: NLNZ - Tutorial D 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:							

SAVE **CANCEL**

4. 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.

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="checkbox"/> http://mp.natlib.govt.nz/	<input type="checkbox"/>	Blanket Authorisation	Dummy Authorising Agency	01/05/2019	30/04/2029	Approved	
<input checked="" type="checkbox"/> https://natlib.govt.nz/	<input checked="" type="checkbox"/>	Blanket Authorisation	Dummy Authorising Agency	01/05/2019	30/04/2029	Approved	

5. 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.

Base Profile
 Harvester Type: HERITRIX3
 Base Profile: Special Profile

Profile Overrides
 Override Imported Profile:

Profile Note:

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

Name	Action
This target does not belong to any groups.	

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

The screenshot shows the 'Targets' interface with a title bar 'NLNZ - Tutorial D'. Below it is a search bar with a placeholder 'Name:' and a 'search' button. The main area has two sections: 'Selected' (which is empty) and 'Results'. The 'Results' section contains a table with columns 'Name', 'Status', and 'Agency'. One row is shown: 'WCT Workshop', 'Active', 'workshop'. Below the table are pagination controls: 'Results 1 to 1 of 1', 'Page | 1 | of 1', and 'Rows per page: 10'. At the bottom are 'SAVE' and 'cancel' buttons, with a green arrow pointing to the 'SAVE' button.

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.

The screenshot shows the 'Target Summary' interface with a title bar 'NLNZ - Tutorial D'. It features a search bar with fields for 'ID', 'From', 'To', 'Agency', and 'Owner'. Below the search are filters for 'Name' (set to 'NLNZ - Tutorial D'), 'Flag' (set to 'None'), 'State' (set to 'Scheduled'), and 'QA Recommendation' (checkboxes for 'Archive', 'Reject', 'Investigate', 'Delist', 'Archived', 'Archiving', and 'Failed'). There are 'search' and 'reset' buttons. The main area is titled 'Results' with a table header: 'Thumbnail', 'Name', 'Harvest Date', 'State', 'Owner', 'Run Time', 'Data Downloaded', 'URLs', '% Failed', 'Crawls', 'QA Recom', and 'Action'. One row is listed: 'NLNZ - Tutorial D', '03/06/2019 09:44:48', 'Running', 'W. User', '00:00:00.01', '0 bytes', '0', '0.0', '1', 'H3'. Below the table are pagination controls: 'Results 1 to 1 of 1', 'Page | 1 | of 1', and 'Rows per page: 10'. A green arrow points to the 'archive' button in the toolbar above the table.

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

The screenshot shows the H3 Script Console interface. A green box highlights the configuration area on the left, which includes fields for Target Instance Oid (1114112), Target Name (NLNZ - Tutorial D), Scripts (list-pending-urls), and Script Engine (Groovy). Below this is the script code area, also highlighted with a green box. The code is a Groovy script that prints pending URLs from a database. At the bottom is an 'Execute Script' button, also highlighted with a green box. Three red arrows point from the text above to these three specific areas: the 'Scripts' dropdown, the 'Script Engine' dropdown, and the 'Execute Script' button.

```
// groovy
// see org.archive.crawler.frontier.BdbMultipleWorkQueues.forAllPendingDo()

import com.sleepycat.je.DatabaseEntry;
import com.sleepycat.je.OperationStatus;

MAX_URLS_TO_LIST = 1000

rawOut.println "Pending URLs:"

pendingUrIs = job.crawlController.frontier.pendingUrIs

cursor = pendingUrIs.pendingUrIsDB.openCursor(null, null);
key = new DatabaseEntry();
value = new DatabaseEntry();
count = 0;
```

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

The screenshot shows the H3 Script Console after executing the 'list-pending-urls' script. A green box highlights the output area, which displays the response code (200), status (0), and a list of pending URLs. The list includes various URLs from the National Library of New Zealand's website, such as eventapi.libring.com/robots.txt, eventapi.libring.com/v1/event, and several accessibility and copyright pages. The output area is also highlighted with a green box. A red arrow points from the text above to the 'Execute Script' button in the previous screenshot, indicating the action that triggered this output.

```
Response Code: 200
Status: 0

Pending URLs:
https://eventapi.libring.com/robots.txt
https://eventapi.libring.com/v1/event
https://images.ctfassets.net/pwv49hug9jad/5G9S7DRLBCC2M8yqoa48kw/a7da447599aa45b4905e63f3d6ea8c33/ATL_033_k
https://images.ctfassets.net/pwv49hug9jad/6MGnvLDcyW0osE6e4s4kEE/ec12e3cab7d4a1f34cb9f3e05037a5e9/feature_c
https://natlib.govt.nz/system/images/WisiziIsIjiwMTkvMDQvMjkvMTZfMTVfMDJfMjYxx0FsYnVtQ292ZXJzXzY2NC5qcGciXv
https://natlib.govt.nz/blog/authors/24
https://natlib.govt.nz/blog
https://natlib.govt.nz/about-us/te-puna-foundation-supporters
https://natlib.govt.nz/about-us/media
https://natlib.govt.nz/about-this-site/contact-us
https://natlib.govt.nz/about-this-site
https://natlib.govt.nz/about-this-site/accessibility-standards
https://natlib.govt.nz/about-this-site/terms-of-use
https://natlib.govt.nz/about-this-site/copyright-and-privacy
https://natlib.govt.nz/about-this-site/help
https://natlib.govt.nz/events/zeros-and-ones-vintage-computers-from-the-national-library-collection-may-20-
https://natlib.govt.nz/events/for-all-the-macs-i-ve-loved-before-may-23-2019
https://natlib.govt.nz/events/navigating-the-complex-world-of-born-digital-collecting-and-preservation-may-
https://natlib.govt.nz/national-library/assets/proxima-nova/36F881_4_0-bb69d2b2194690160b67f233adcd30df751t
https://natlib.govt.nz/national-library/assets/proxima-nova/36F881_5_0-b4c5bf0695abcdacee3573570bb57af2d&
https://natlib.govt.nz/national-library/assets/proxima-nova/36F881_6_0-220c9a4cb2ff2a58a1dd85135aa100d8ba6t
https://natlib.govt.nz/national-library/assets/proxima-nova/36F881_9_0-a06de80ebb14335e50020e1e184d2b7b7f1c
```

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.

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

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.

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

7. Click the *Tree View* link

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

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.

Target Summary
NLNZ - Tutorial D (1114112)

Resource

- Harvest
 - http://cdn.embedly.com/
 - http://embed.ly/
 - http://jquery.com/
 - http://leefletjs.com/
 - http://mp.natlib.govt.nz/?=en
 - http://mp.natlib.govt.nz/?=mi
 - http://natlib-primo.hosted.exlibrisgroup.com/
 - http://natlib.govt.nz/
 - http://natlib.govt.nz/kakapo/
 - http://natlib.govt.nz/national-library/
 - http://natlib.govt.nz/national-library/assets/
 - http://natlib.govt.nz/b00051a2dacbe60c9f26d80028356c09f7d7aaea6713c270b3ea588ee7981a6c.png
 - http://natlib.govt.nz/robots.txt
 - http://natlib.govt.nz/custom/

Viewing

Pruning

Importing
All URLs must be of the form `http(s)://[host-name]/[file-path]/[file-name]` e.g <http://crawledsite.com/images/logo.jpg>

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.

Target Summary
NLNZ - Tutorial D (1114112)

Resource

- http://natlib.govt.nz/national-library/
 - http://natlib.govt.nz/national-library/assets/
 - http://natlib.govt.nz/national-library/assets/logo-b00051a2dacbe60c9f26d80028356c09f7d7aaea6713c270b3ea588ee7981a6c.png
 - http://natlib.govt.nz/records/
 - http://natlib.govt.nz/records/
 - http://natlib.govt.nz/robots.txt
 - http://natlib.govt.nz/system/
 - http://nhaddeliver.natlib.govt.nz/
 - http://search.ebscohost.com/
 - http://taki.natlib.govt.nz/
 - http://vimeo.com/
 - http://www.w3.org/
 - http://www.youtube.com/
 - http://www.youtube.com/robots.txt
 - http://www.youtube.com/watch?v=
 - https://api.cdn.embed.ly/
 - https://apis.google.com/
 - https://app.embed.ly/
 - https://askalibrarian.natlib.govt.nz/

Viewing

Pruning

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.

Import from disk file
Source File: Choose File No file chosen

Import from URL
Source URL: Use Selected

Automated QA Imports
Url
Select all | De-select all
There are no URLs available for import

Provenance Note:
Removing unwanted content from crawl

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.

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	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	

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

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	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

Tutorial E - User Management

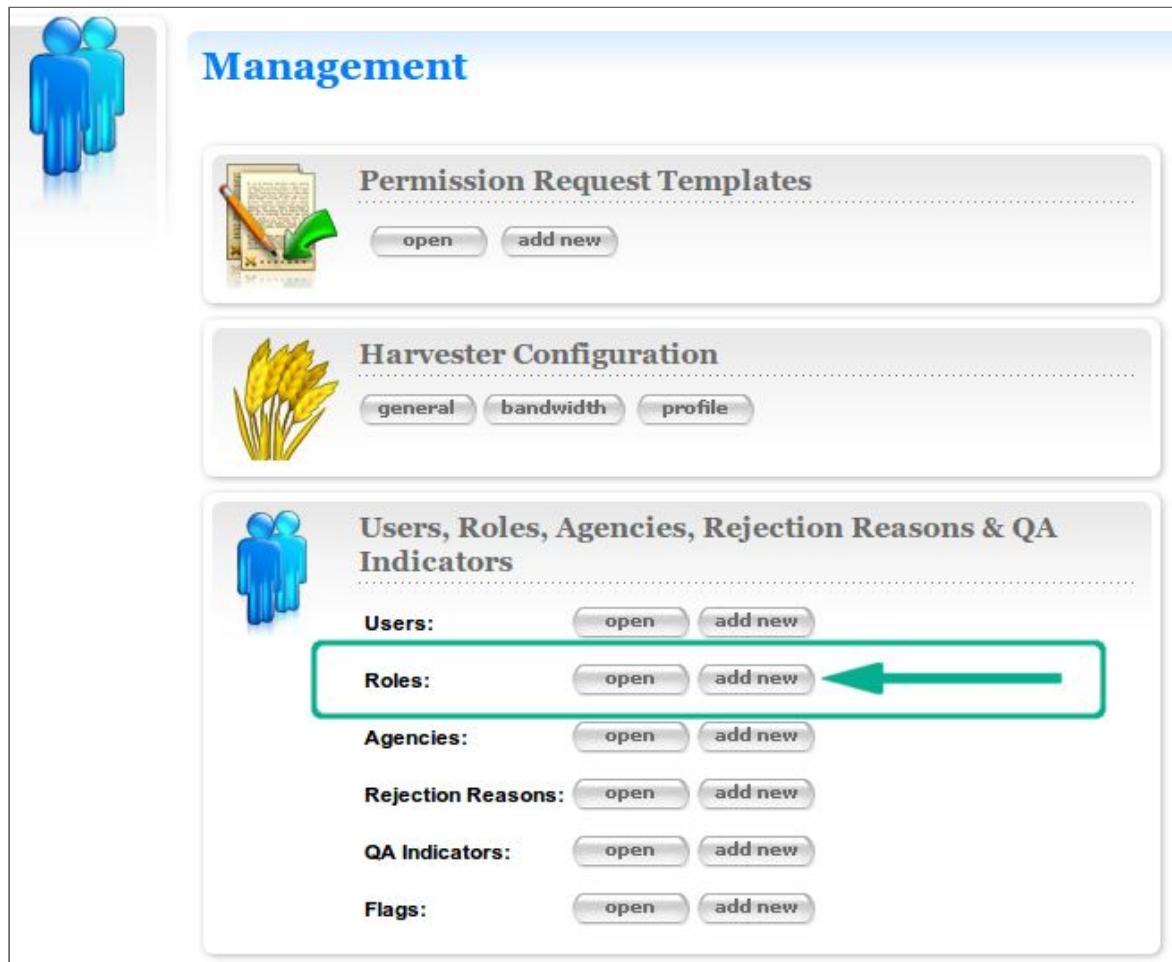
What you will learn [user management]

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.



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

Manage Copying Permissions and Access Rights

<input type="checkbox"/> Create Harvest Authorisations	<input type="button" value="Agency ▼"/>
<input type="checkbox"/> Modify Harvest Authorisations	<input type="button" value="All ▼"/>
<input type="checkbox"/> Confirm Permissions	<input type="button" value="Agency ▼"/>

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

<input type="checkbox"/> Set Harvest Profile Level 1	<input type="button" value="Agency ▼"/>
<input type="checkbox"/> Set Harvest Profile Level 2	<input type="button" value="Agency ▼"/>
<input type="checkbox"/> Set Harvest Profile Level 3	<input type="button" value="Agency ▼"/>

Manage Harvests

<input checked="" type="checkbox"/> Manage Target Instances	<input type="button" value="Agency ▼"/>
<input checked="" type="checkbox"/> Launch Target Instance Immediate	<input type="button" value="Agency ▼"/>
<input checked="" type="checkbox"/> Manage Web Harvester System	<input type="button" value="All ▼"/>
<input checked="" type="checkbox"/> Endorse Harvest	<input type="button" value="Agency ▼"/>
<input checked="" type="checkbox"/> UnEndorse Harvest	<input type="button" value="Agency ▼"/>
<input checked="" type="checkbox"/> Archive Harvest	<input type="button" value="Agency ▼"/>

Manage Target Groups

<input type="checkbox"/> Create Target Group	<input type="button" value="Agency ▼"/>
<input type="checkbox"/> Add Target to Group	<input type="button" value="Owner ▼"/>

7. Save the role.

Ownership

Give Ownership to

Take Ownership from

Agency ▾

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.

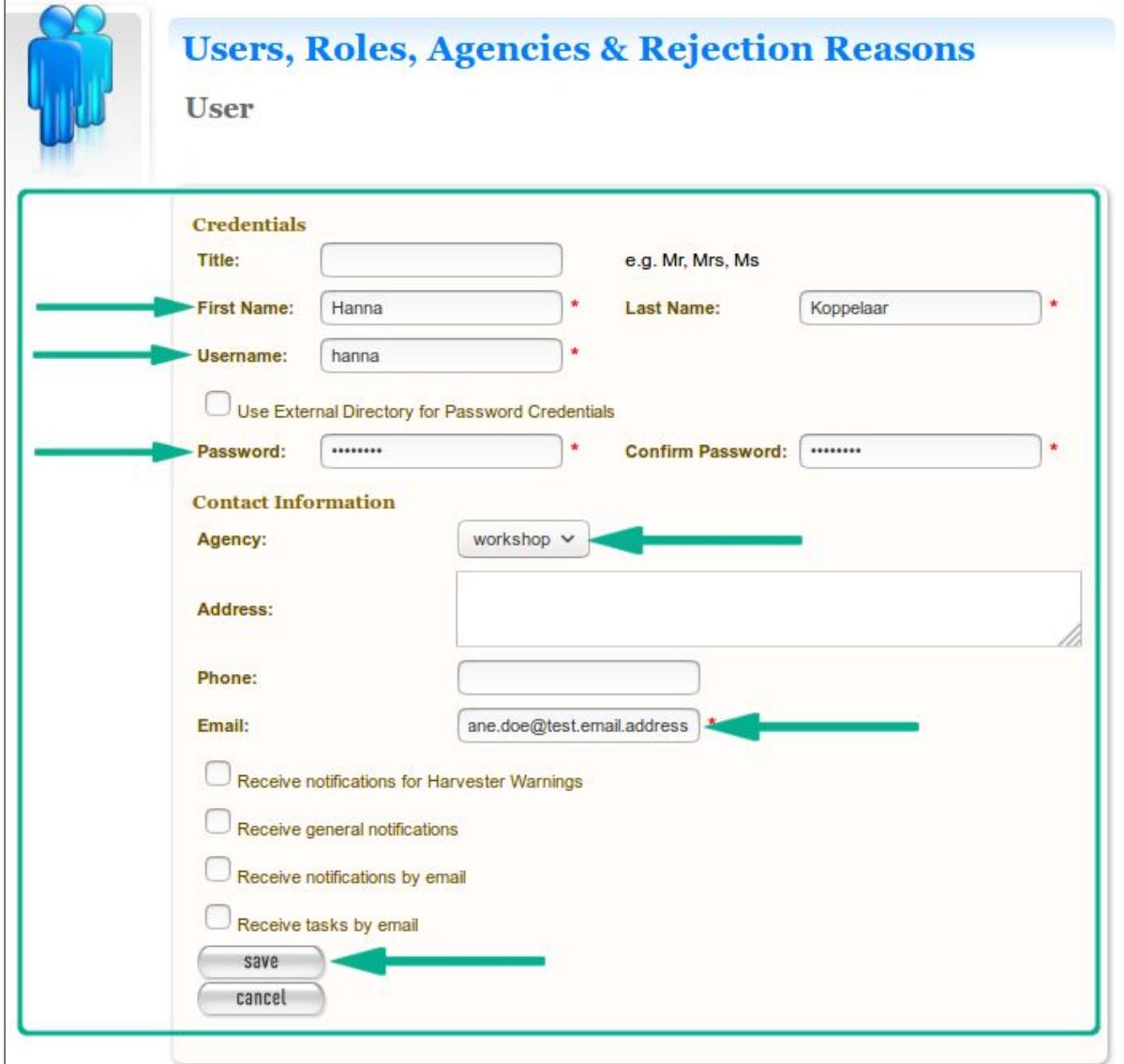
3. Enter and confirm a password.

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

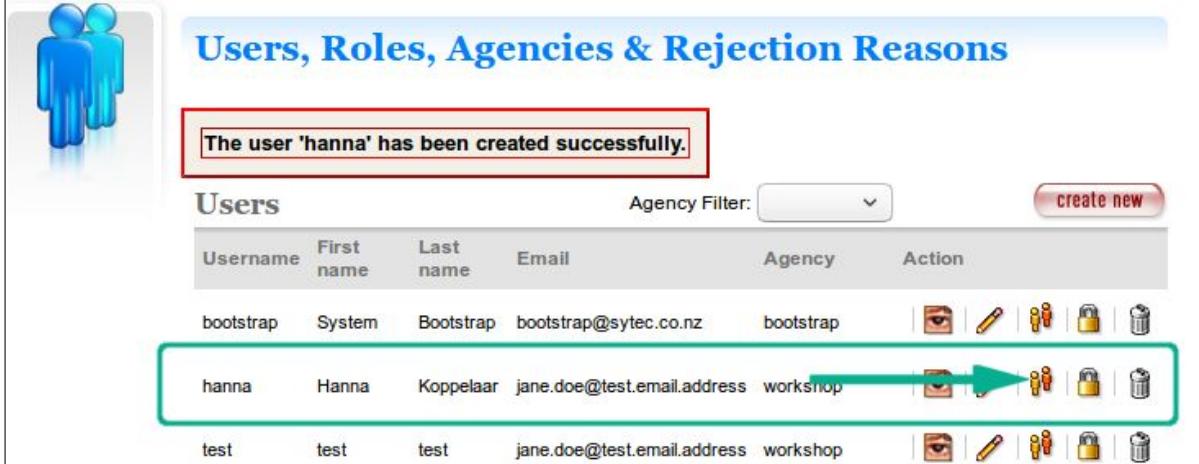
Receive general notifications

Receive notifications by email

Receive tasks by email



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



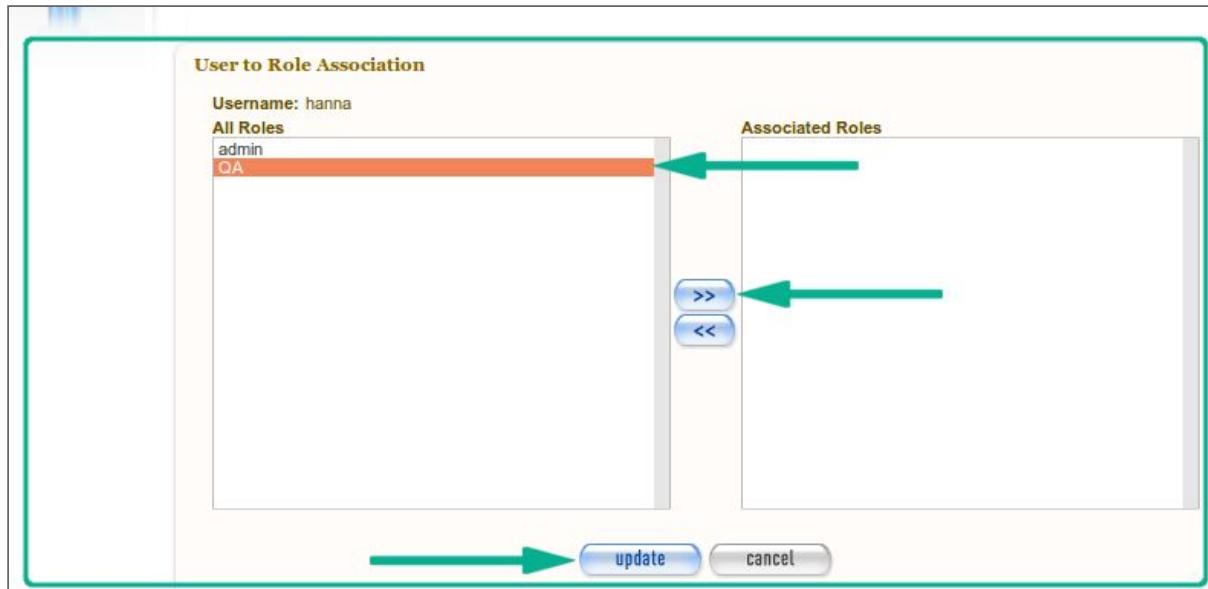
Users, Roles, Agencies & Rejection Reasons

The user 'hanna' has been created successfully.

Users						Agency Filter:	<input type="button" value="create new"/>
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.

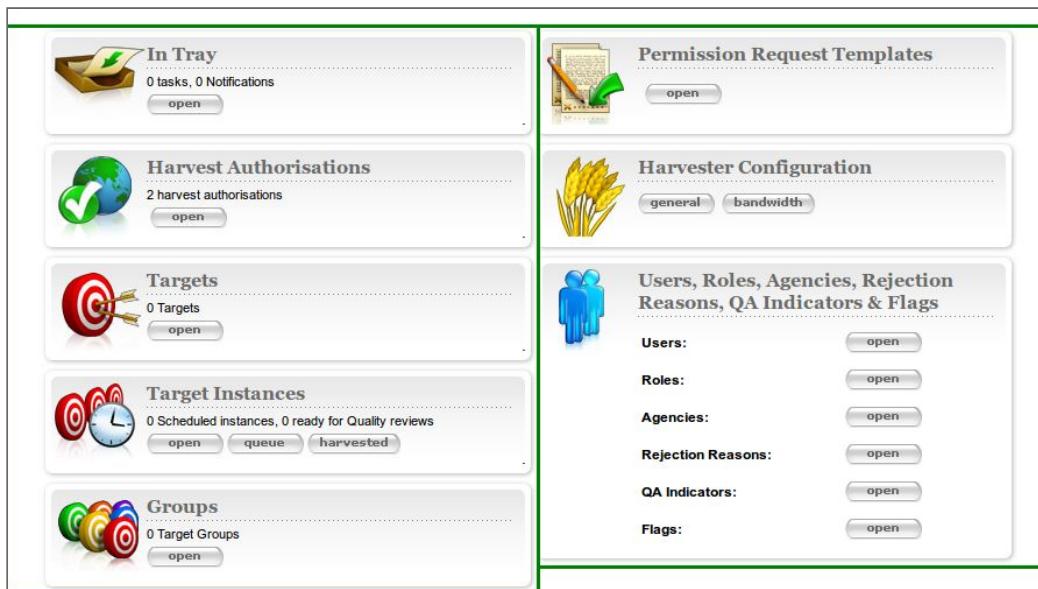
9. Click on the update button.



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.

The screenshot shows the 'Management' interface with several tabs:

- Permission Request Templates:** Shows an icon of a document and a pencil, with 'open' and 'add new' buttons.
- Harvester Configuration:** Shows an icon of wheat, with 'general', 'bandwidth', and 'profile' buttons.
- Users, Roles, Agencies, Rejection Reasons & QA Indicators:** Shows icons for users, roles, agencies, rejection reasons, and QA indicators, each with 'open' and 'add new' buttons.
- Reports:** Shows an icon of a pie chart and a folder, with an 'open' button. A green arrow points to this 'open' button.

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

The screenshot shows the 'Reports' interface with a list of reports:

Title	Description	Action
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

3. In Start Date enter the first date of last month, in End Date enter the date of tomorrow.
4. 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: 01/05/2019 * dd/MM/yyyy

End Date: 01/06/2019 * dd/MM/yyyy

Agencies: All agencies (Optional)

Users: All users (Optional)

generate **cancel**

5. 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

print **save** **e-mail** **cancel**