Web Curator Tool Lesson Plan

IIPC 2019 - WCT Workshop - Hands-on Session

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.

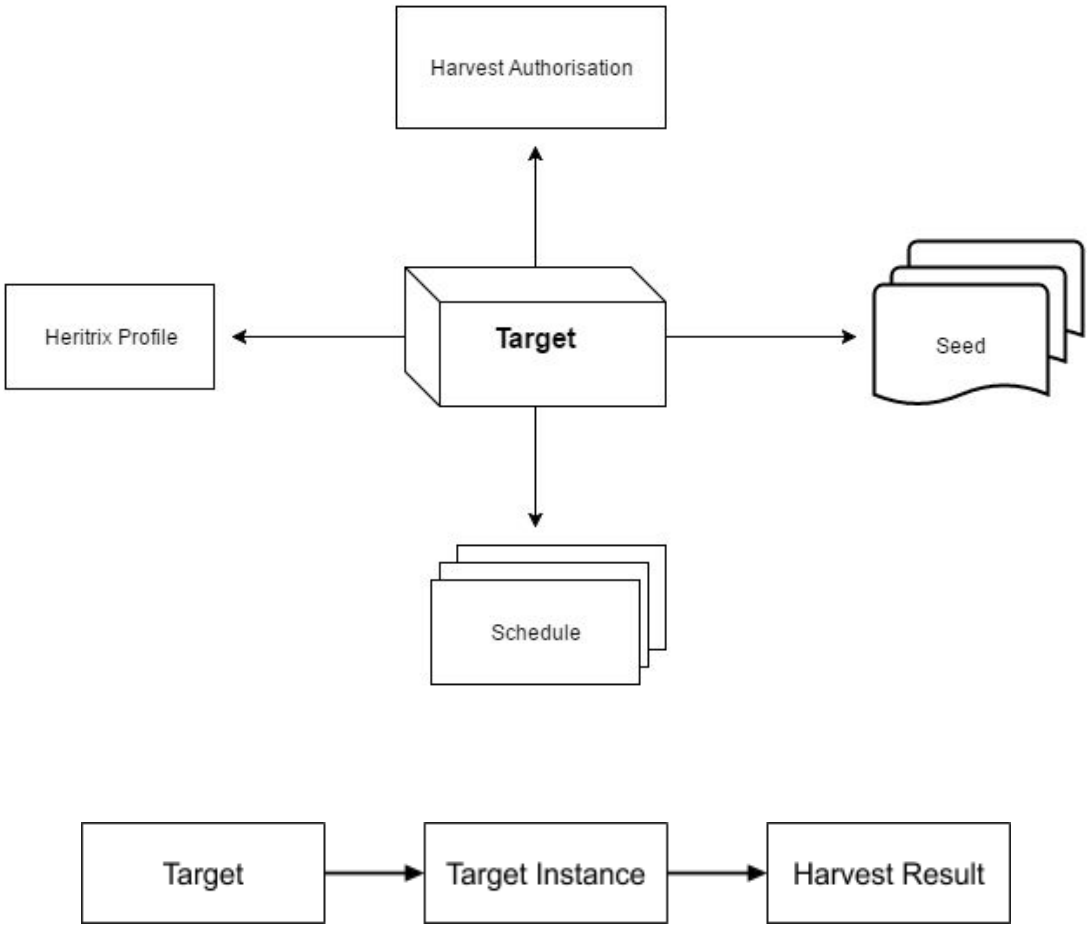
This lesson plan was written by the National Library of New Zealand and National Library of the Netherlands for the WCT Workshop held at the 2019 IIPC General Assembly, based on version 2.x of WCT.

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

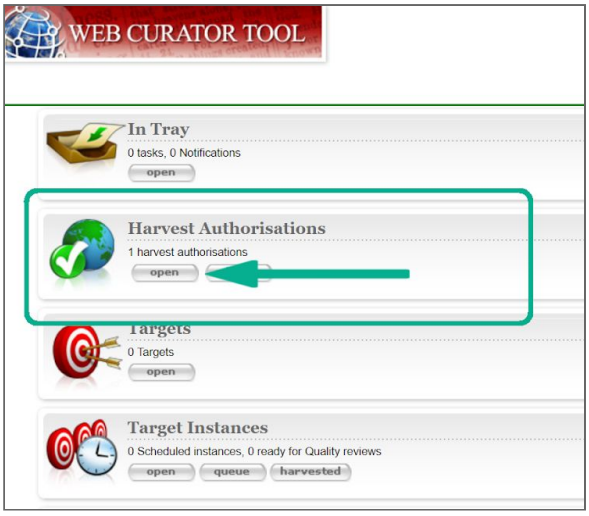
What you will learn [authorisations, profiles, scheduling, crawl monitoring, quality review]

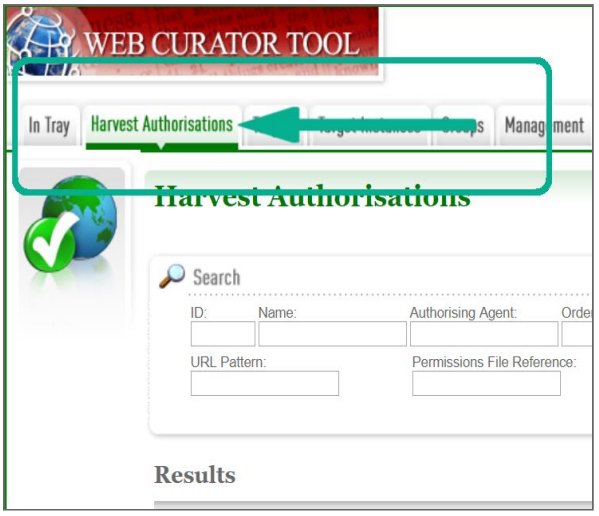
Scenario: You must crawl a website that your institution has determined requires permission to harvest.

Setup Harvest Authorisation

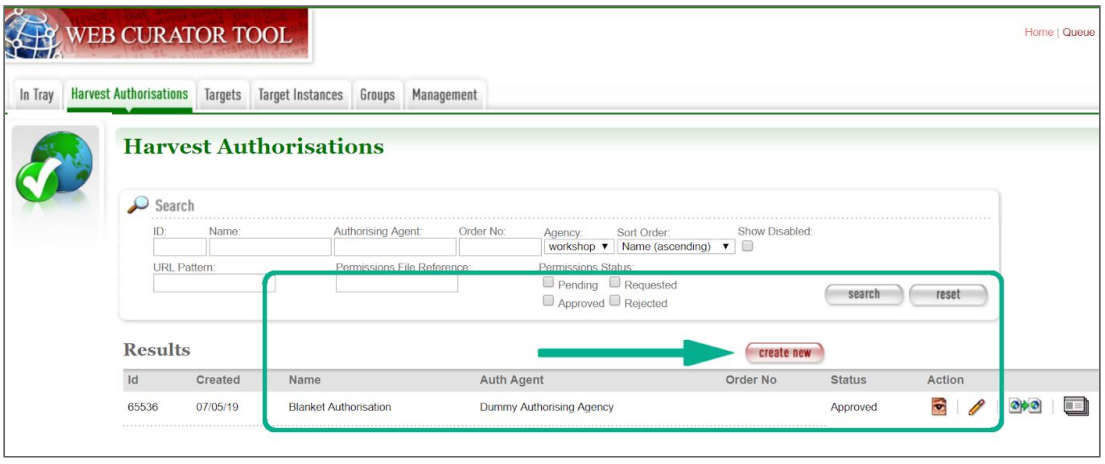
A Harvest Authorisation is a record of requested permission from copyright holders (e.g. website owner) to harvest a website. You must consider copyright law when you harvest material, and also when you preserve it and when you make it accessible to users.
For more information on Harvest Authorisations in WCT, see the documentation.

1. Navigate to the Harvest Authorisation tab in WCT

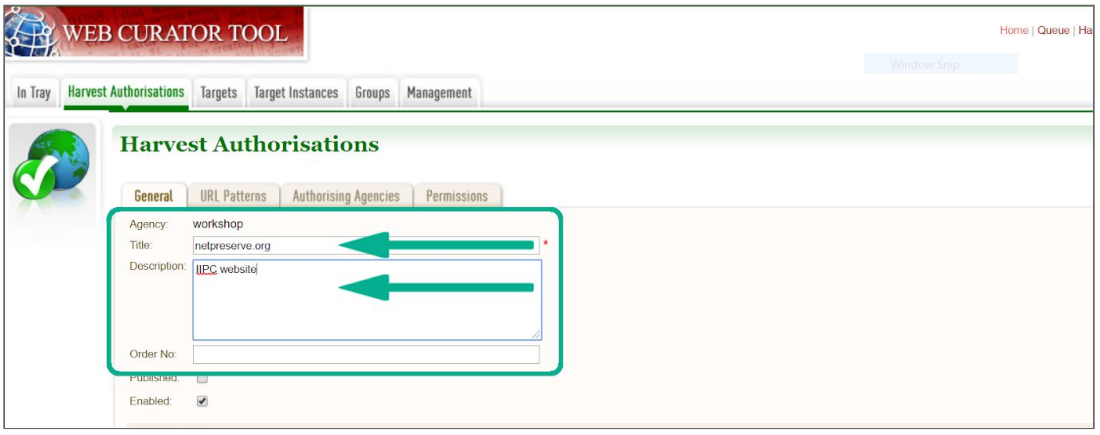




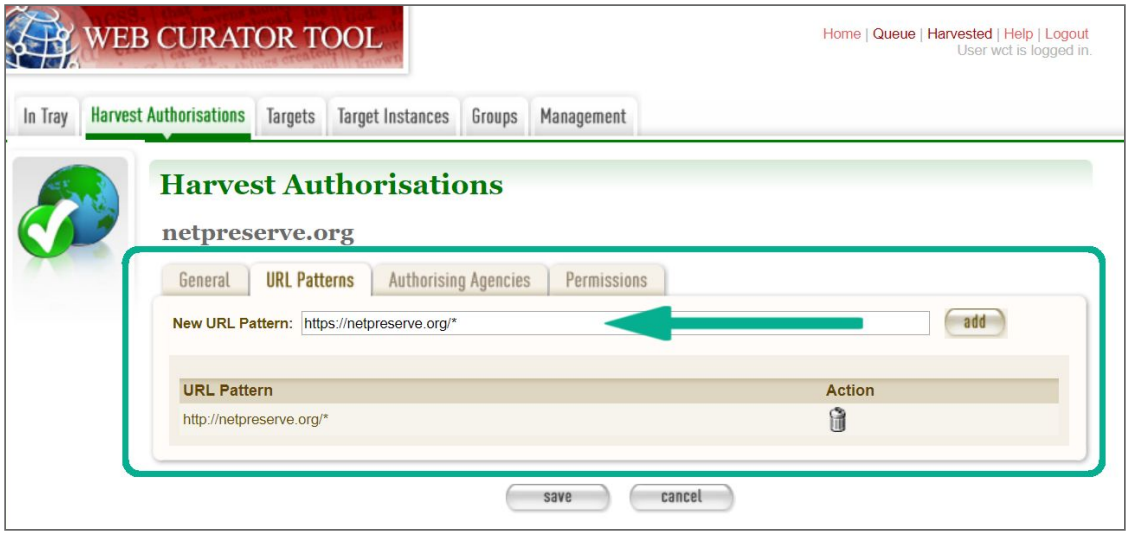
2. Create a new Harvest Authorisation



3. In the General tab enter the following detail
   Title: netpreserv.org
   Description: IIPC Website



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/**

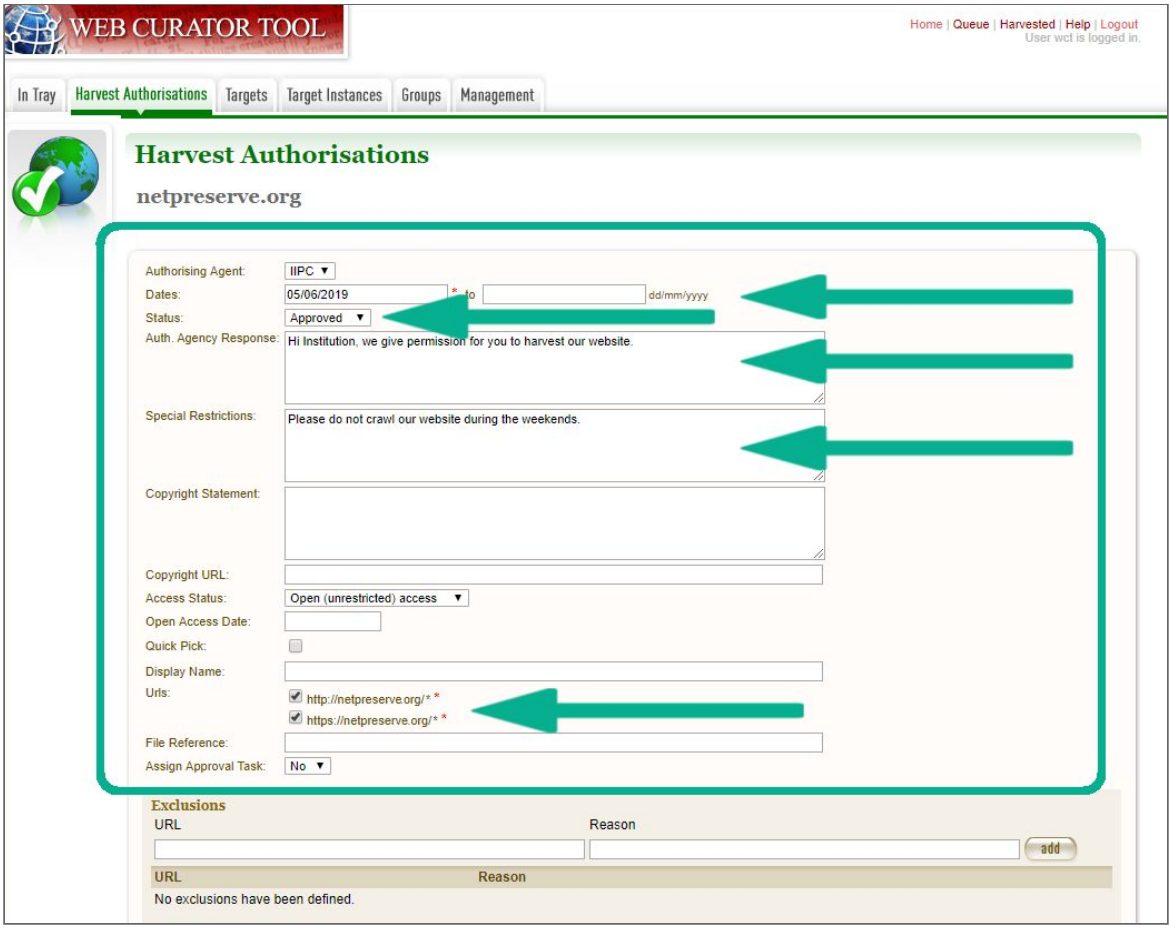


5. Navigate to the Authorising Agencies tab, and click Create New.

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

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

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>

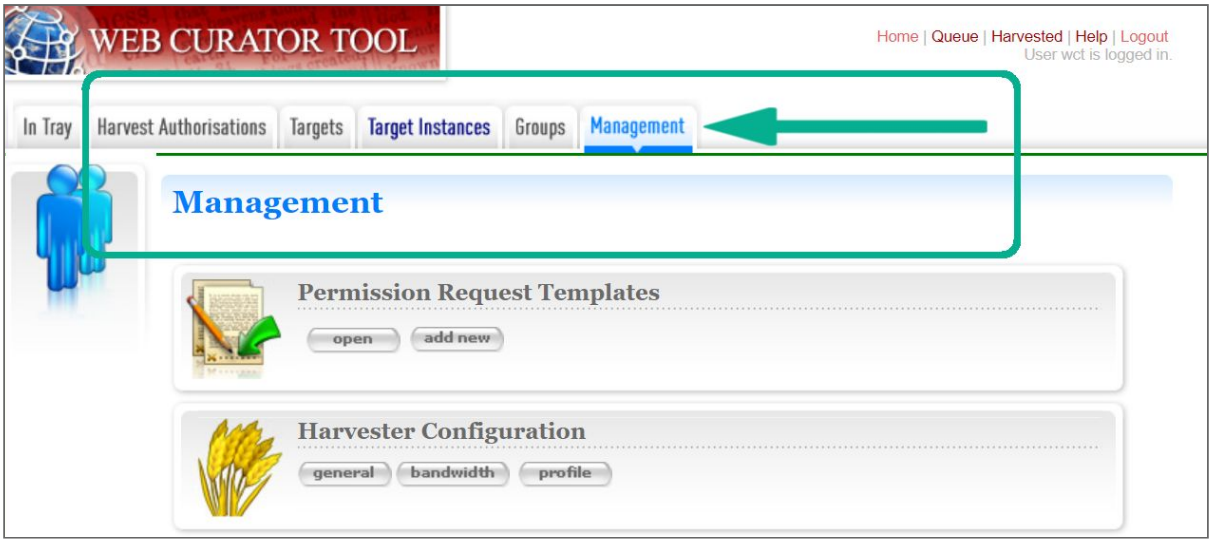


9. Save the new Harvest Authorisation.

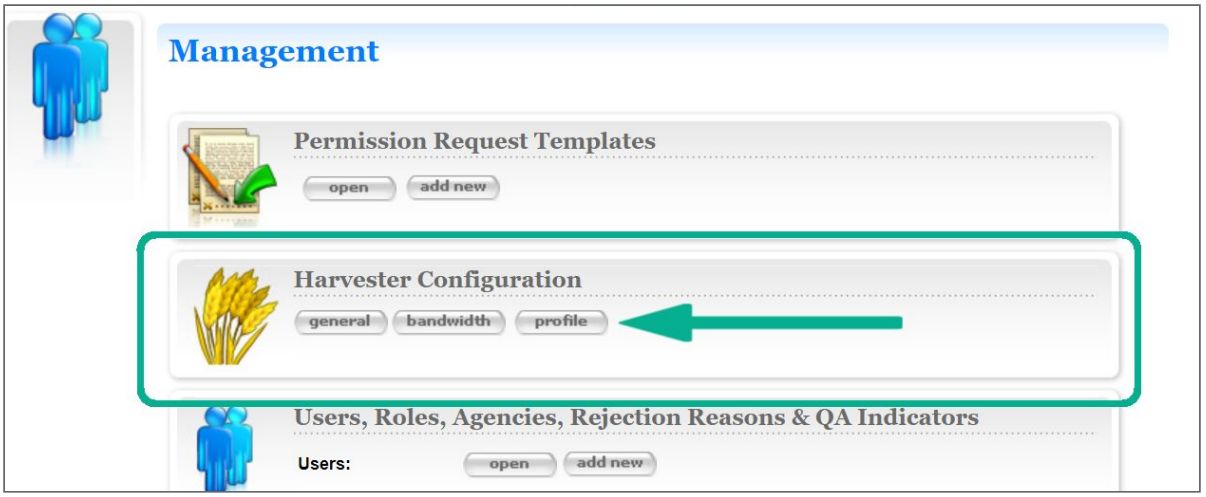
Configure default profile

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.

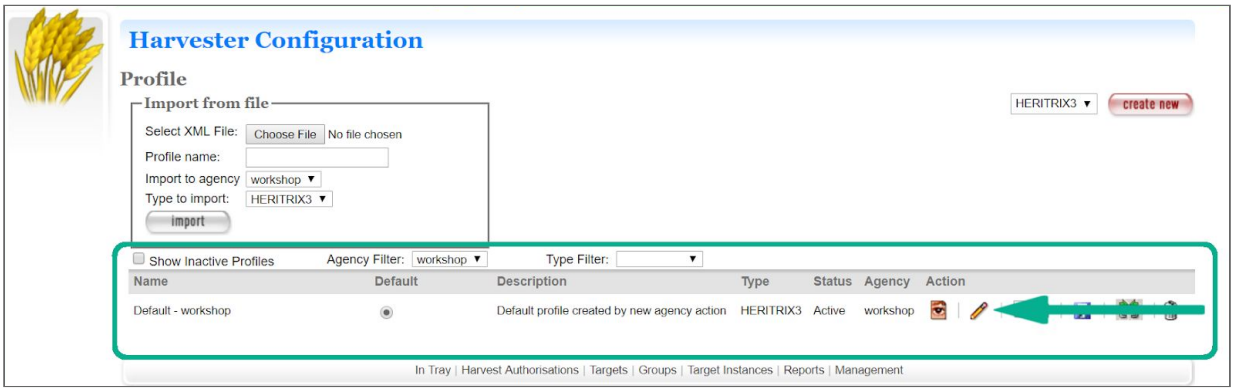
1. Navigate to the Management tab in WCT



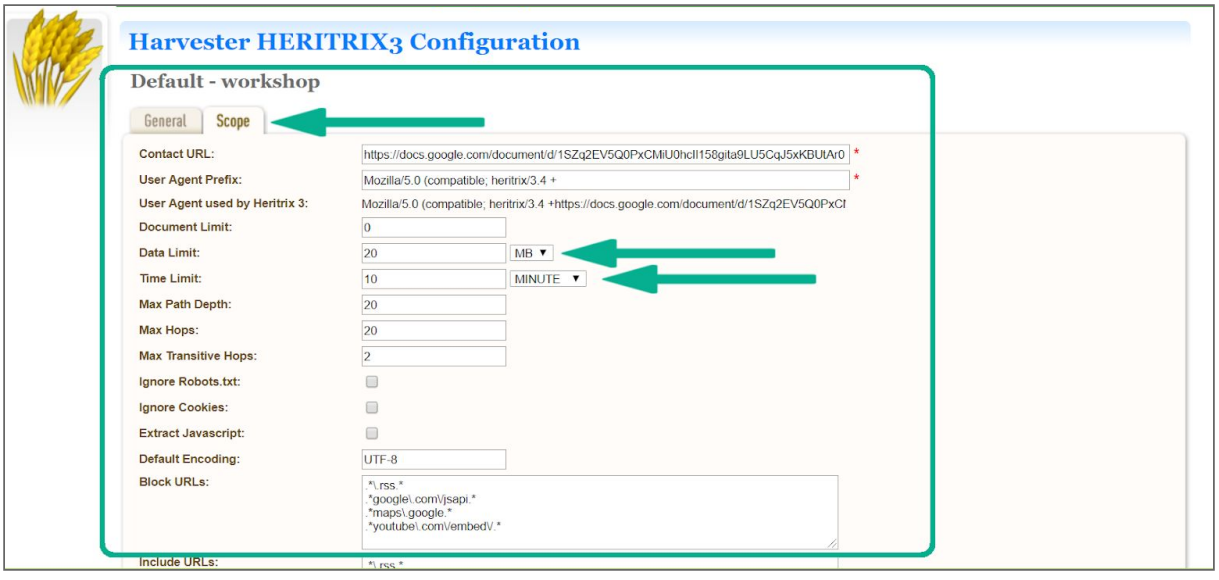
2. Click the Profile option under Harvester Configuration.



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



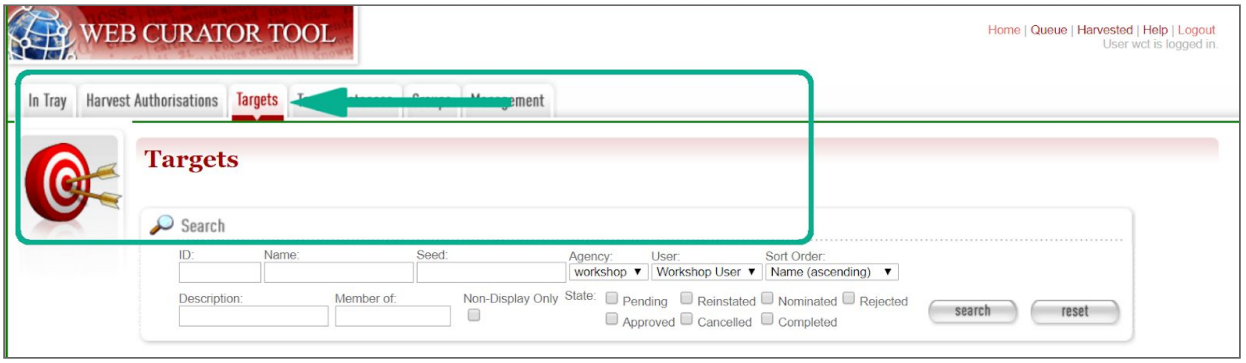
4. Navigate to the Scope tab. We will set data and time limits in the profile, so we don’t wait too long to continue this tutorial. Enter the following settings and click Save.
   Data Limit: 20 MB
   Time Limit: 10 Minute



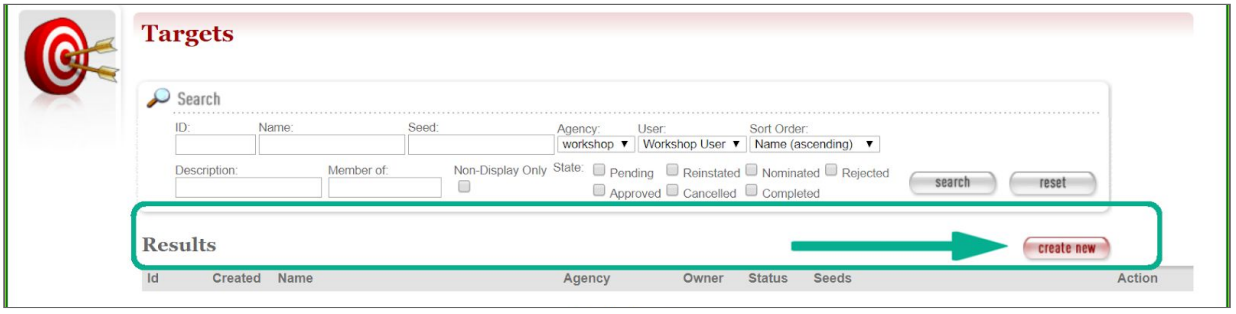
Setup Target

A Target is a record that describes the thing you intend to harvest from the Web. It contains metadata descriptions, seed URLs, how and when the thing will be crawled.
For more information on WCT Targets, see the documentation.

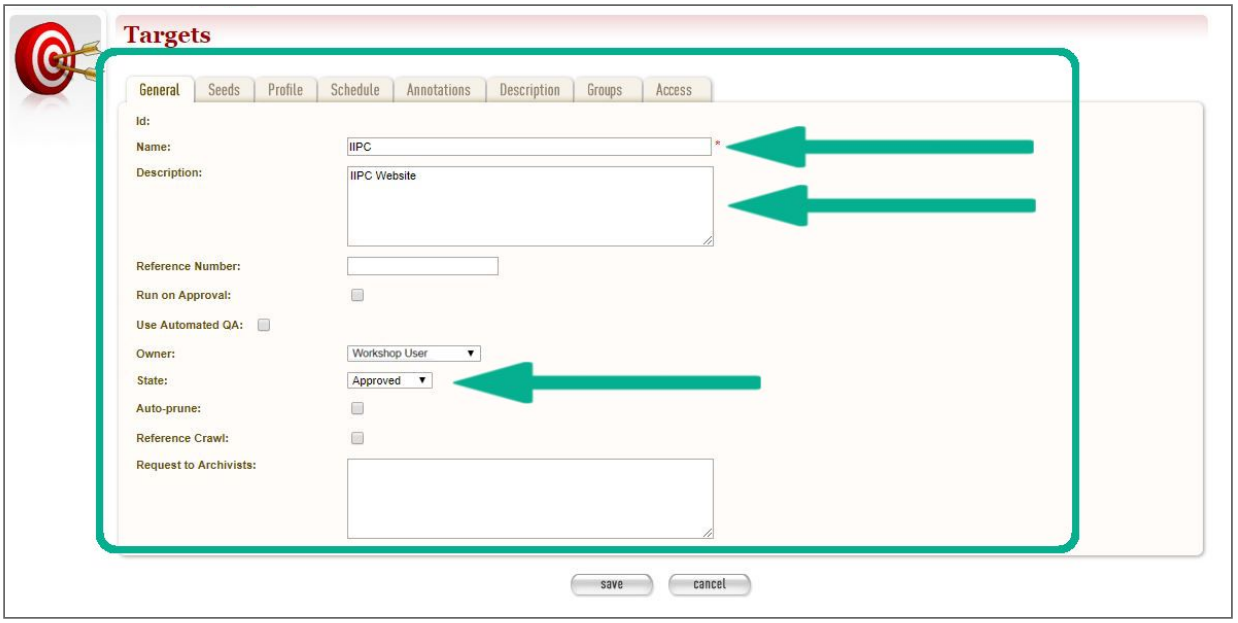
1. Navigate to the Targets tab in WCT



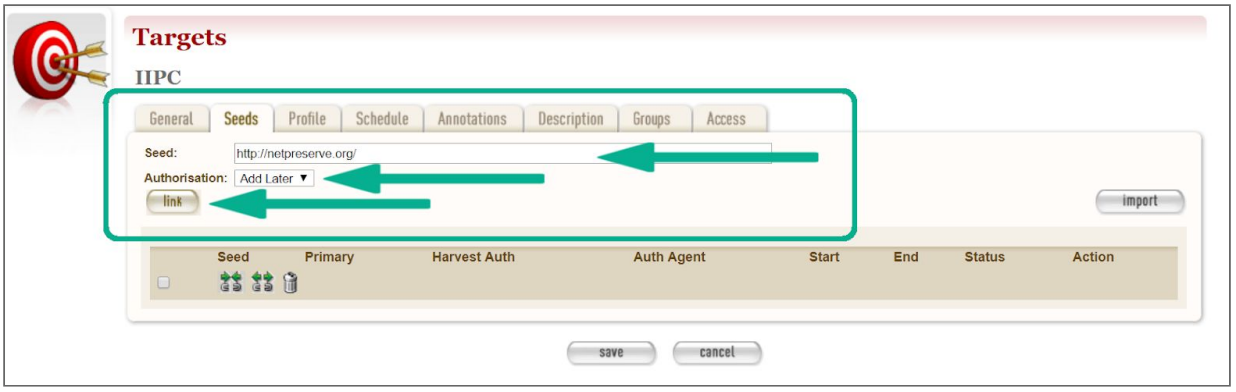
2. Create a new Target



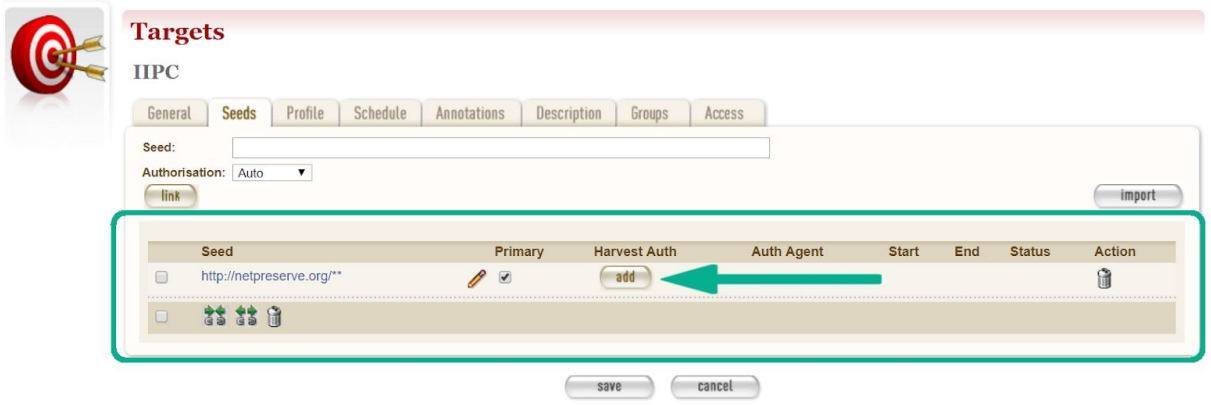
3. In the General tab, enter the following detail
   Name: IIPC
   Description: IIPC Website
   State: Approved



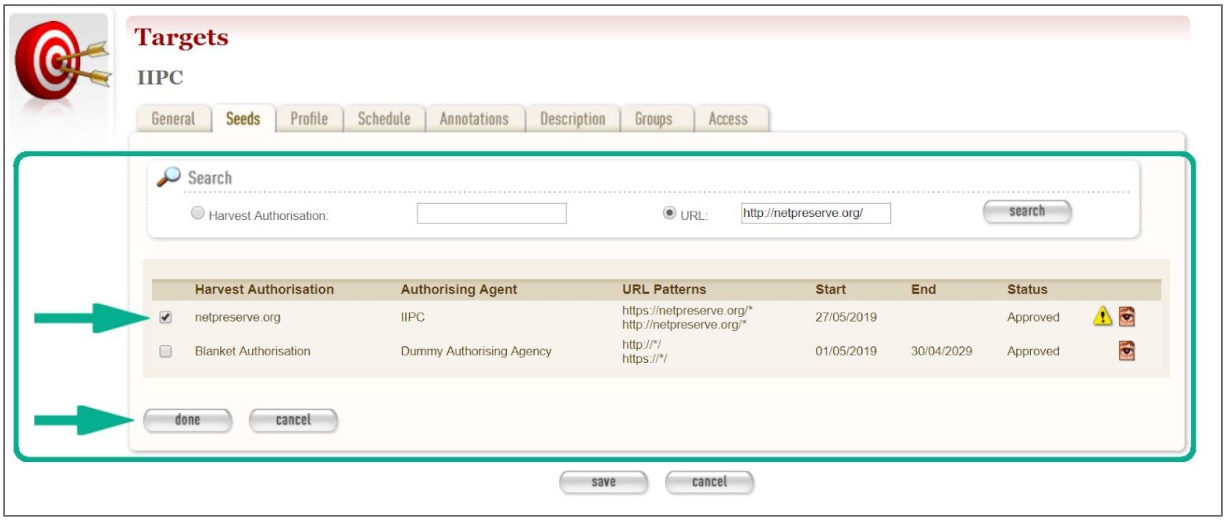
4. Navigate to the Seeds tab, and enter the following detail, then click the Link button.
   Seed: http://netpreserve.org/
   Authorisation: Add Later



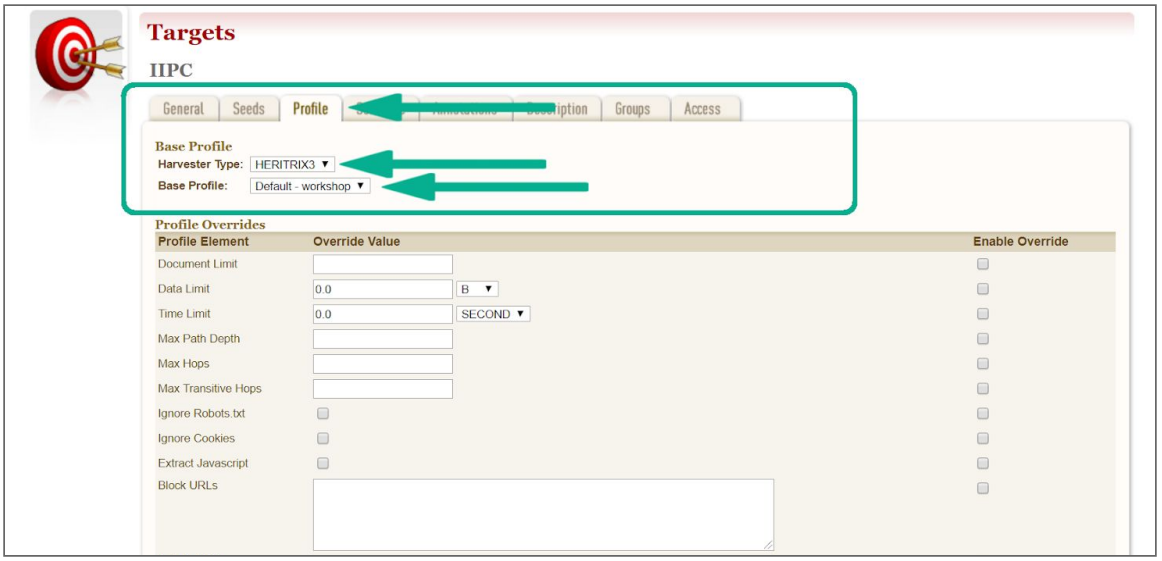
5. Click the Add button to assign a Harvest Authorisation to this seed.
   Each seed must be linked to an authorisation which details the permission granted to harvest content from that URL.



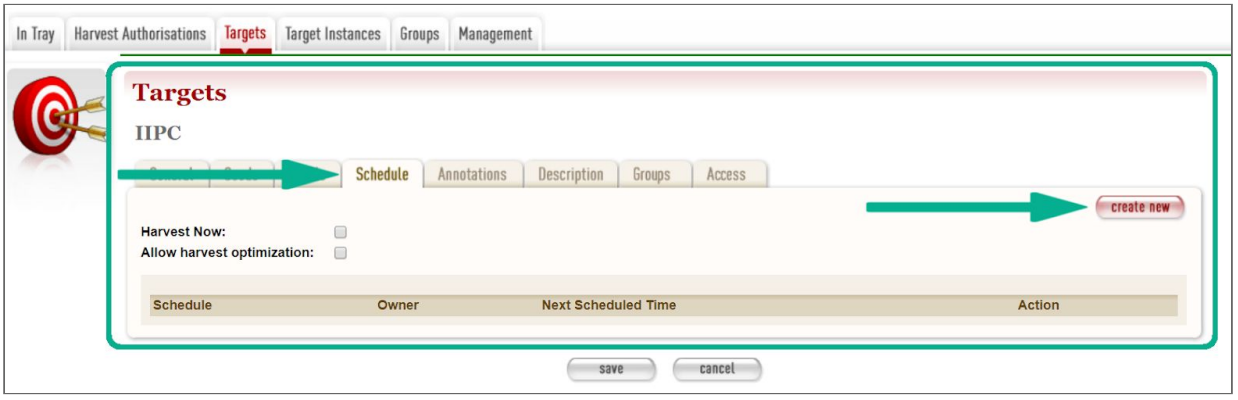
6. Tick the box next to the netpreserve.org Harvest Authorisation and click Done.
   Notice the yellow warning symbol, indicating that the Permission has special restrictions. Click the View icon next to it, to see the Permission details.



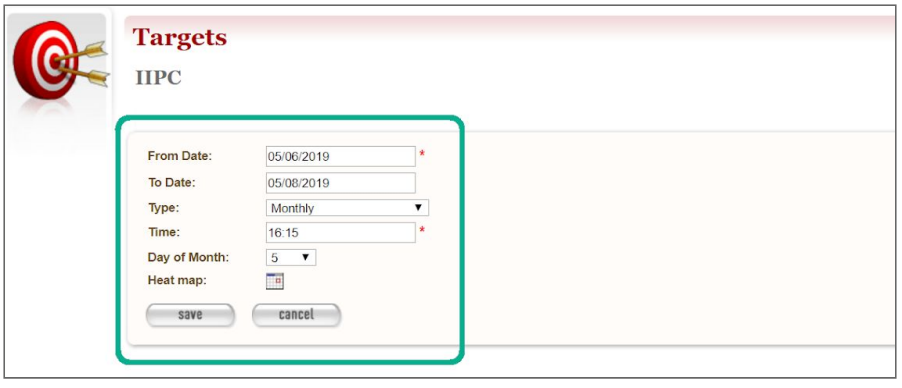
7. Navigate to the Profile tab, and assign the profile to be used during the crawl. Ensure the following detail is set.
   Harvester Type: HERITRIX3
   Base Profile: Default - workshop



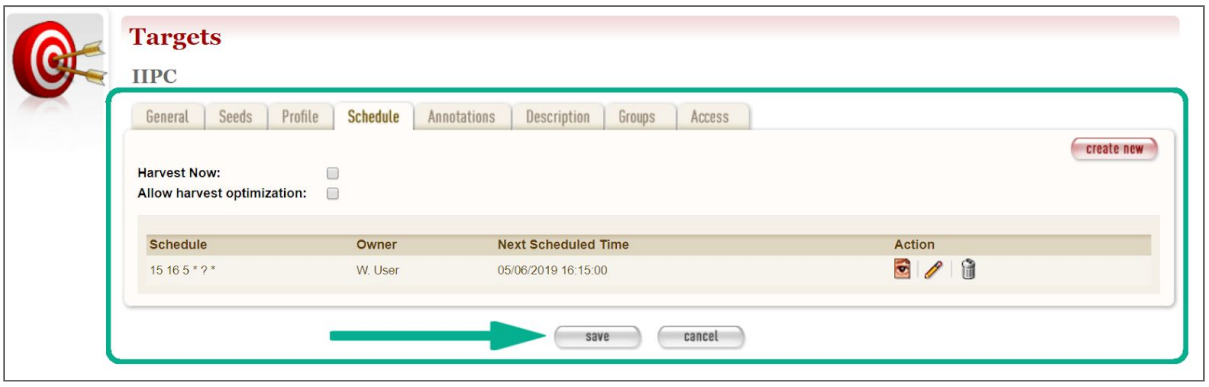
8. Navigate to the Schedule tab, click Create New.



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>



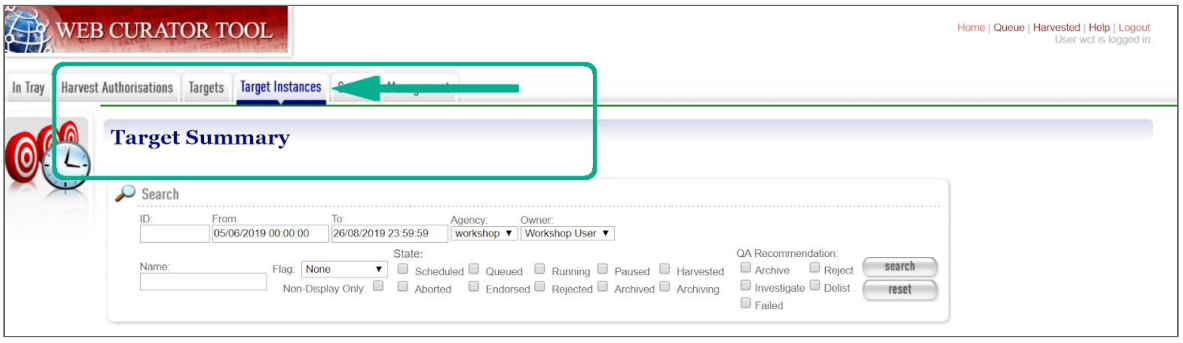
10. Save the Target. The future crawls for this Target will now be scheduled.



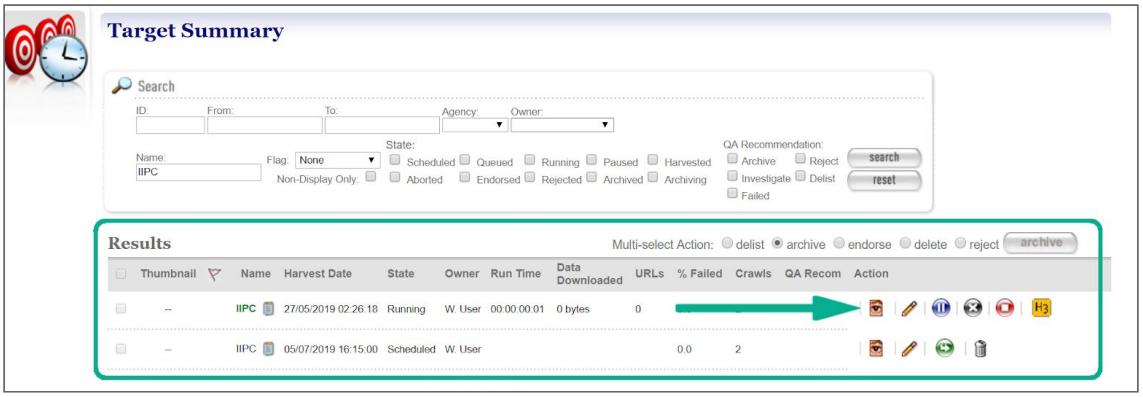
Monitor the Target Instance

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

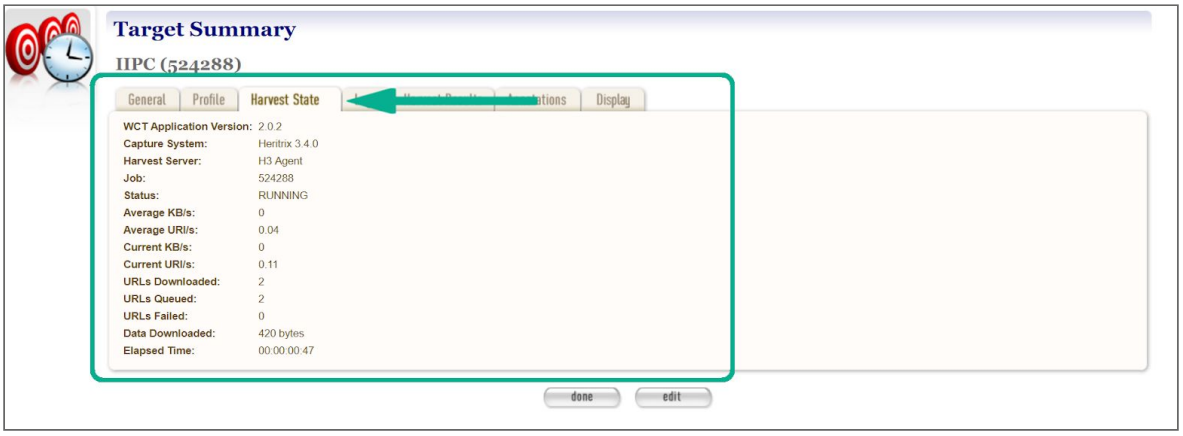
1. Navigate to the Target Instance tab



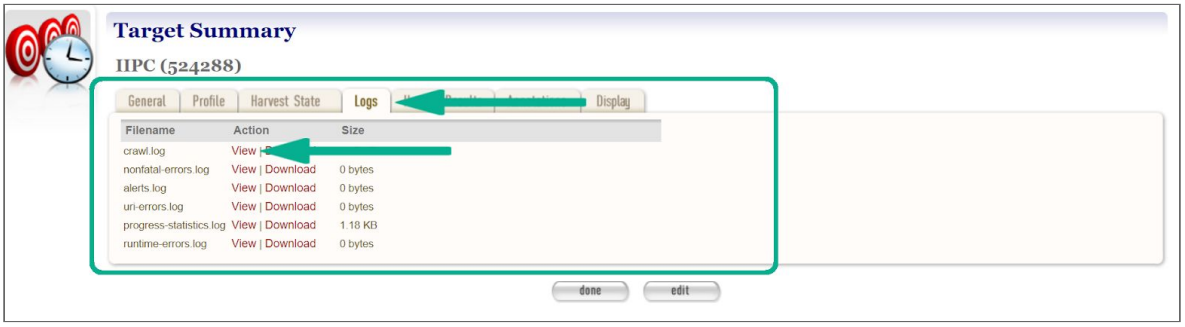
2. The Target we saved earlier should have scheduled a new Target Instance to start five minutes into the future. Observe the scheduled Target Instance move into a Running state. Click the View icon.



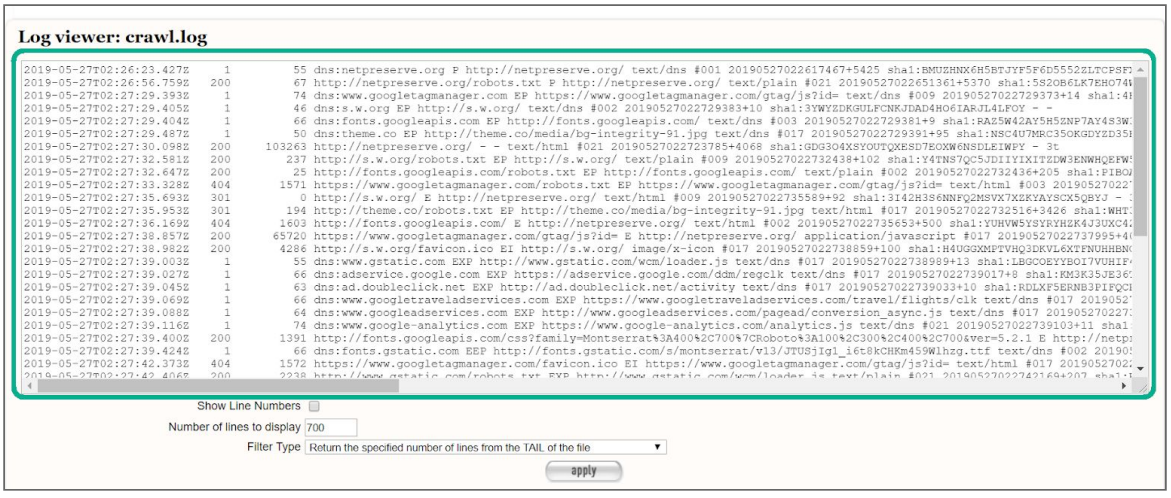
3. Navigate to the Harvest State tab, and observe the crawl statistics.
   You can see:
   - *how many URLs have been downloaded so far, and how many are still in the queue.*
   - *the current and average data and URLs downloaded per second.*
   - *the total data downloaded and crawl time elapsed.*



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



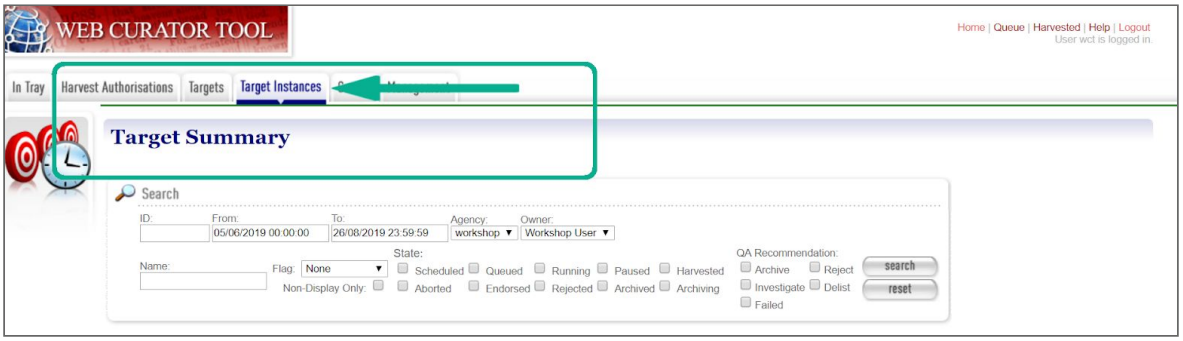
5. Observe the output from the Heritrix crawl log.
   You can see all the URLs that Heritrix has attempted to download, successfully or otherwise.



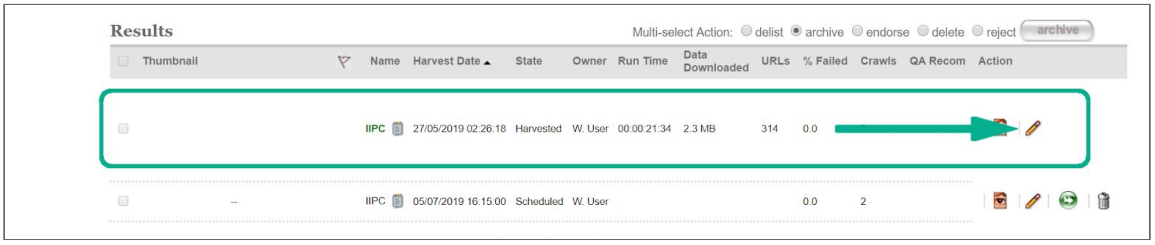
Quality review

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

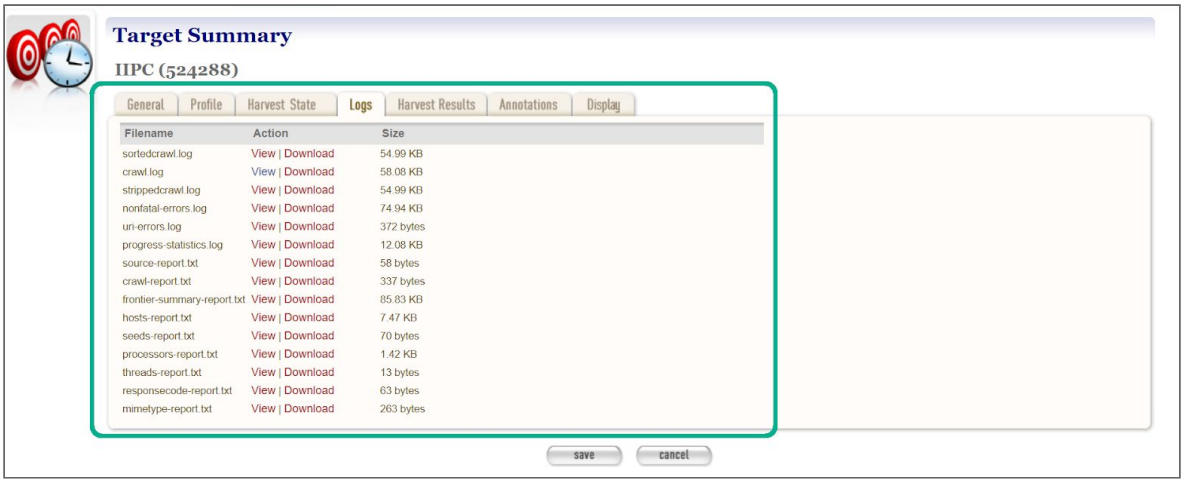
1. Navigate to the Target Instance tab



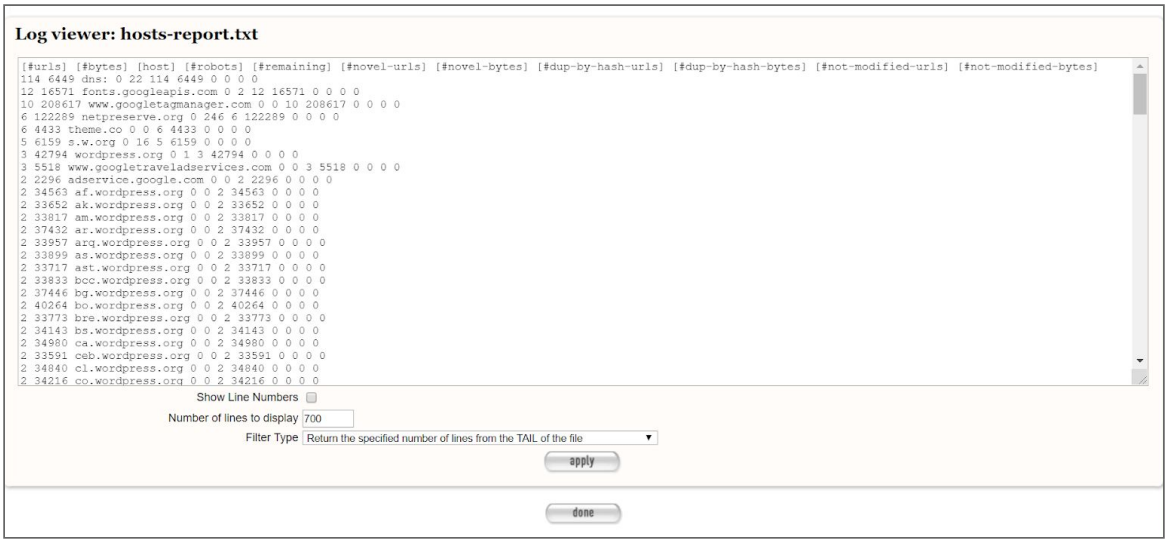
2. When the running Target Instance has finished and moved into the Harvested state, click the Edit icon.



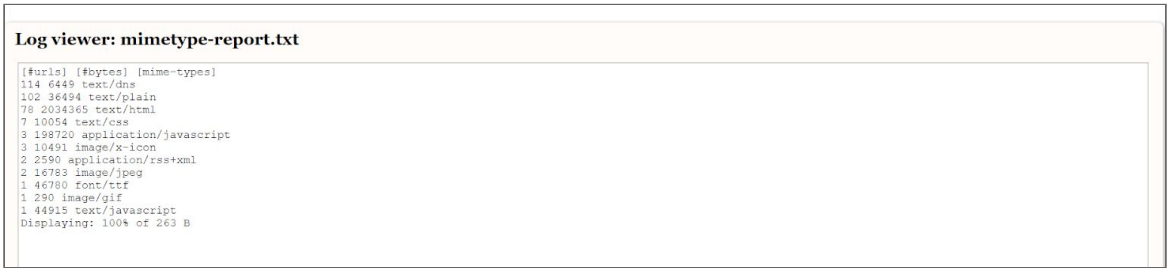
3. Navigate to the Logs tab, and observe the now larger list of available Heritrix logs. These additional reports are generated after a Target Instance is completed.



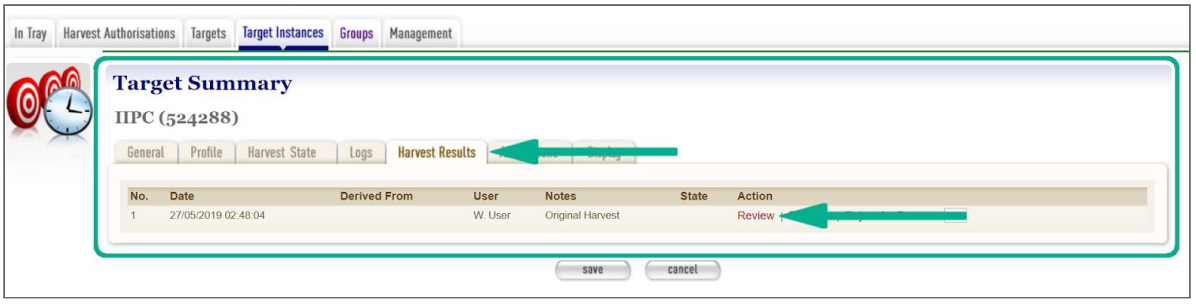
4. The hosts-report will show a breakdown of the number of URLs and bytes crawled per hostname. Once you are finished viewing a log or report, click the Done button.



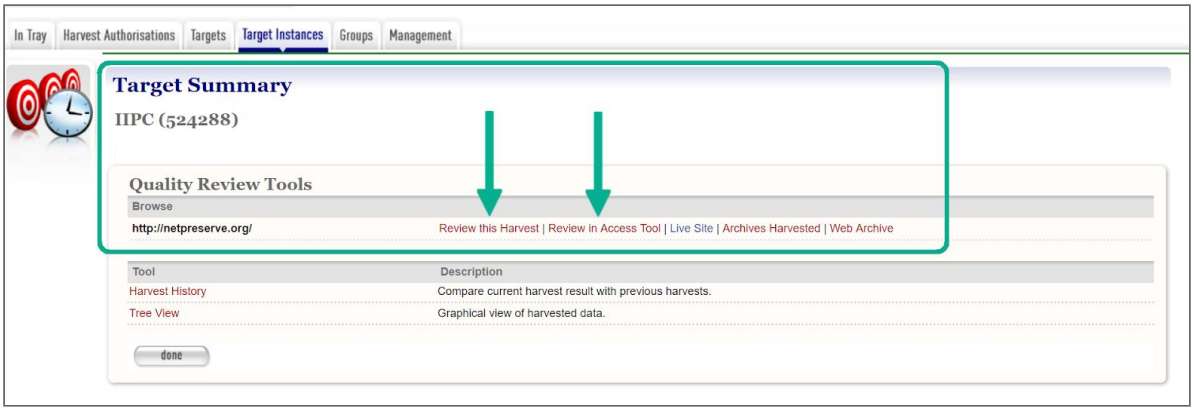
5. The mimetype-report will show a breakdown of the number of URLs and bytes crawled per mime type.



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.



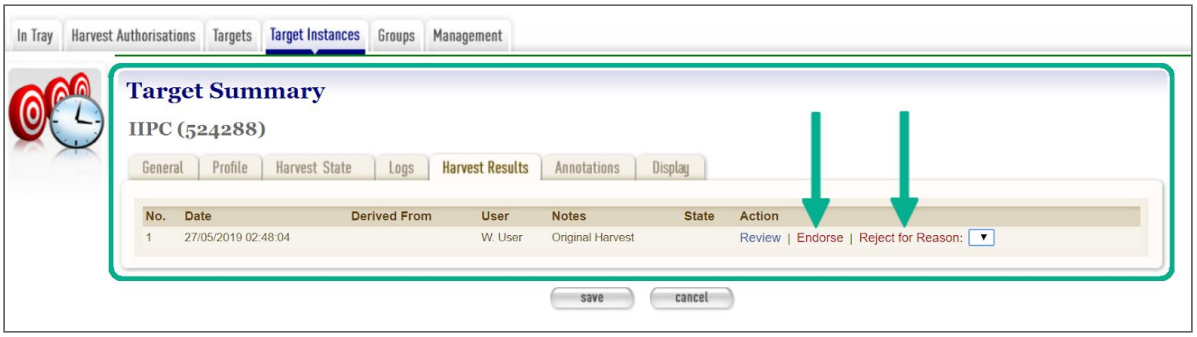
7. There are two options for viewing the harvested result. The ‘Review this Harvest’ link will open the harvest in an old built-in web harvest viewer. The ‘Review in Access Tool’ link will open the harvest in a configured instance of OpenWayback. Click the Review in Access Tool link.



8. You can browse and review the harvest here.

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.



Tutorial B - Crawl Scoping

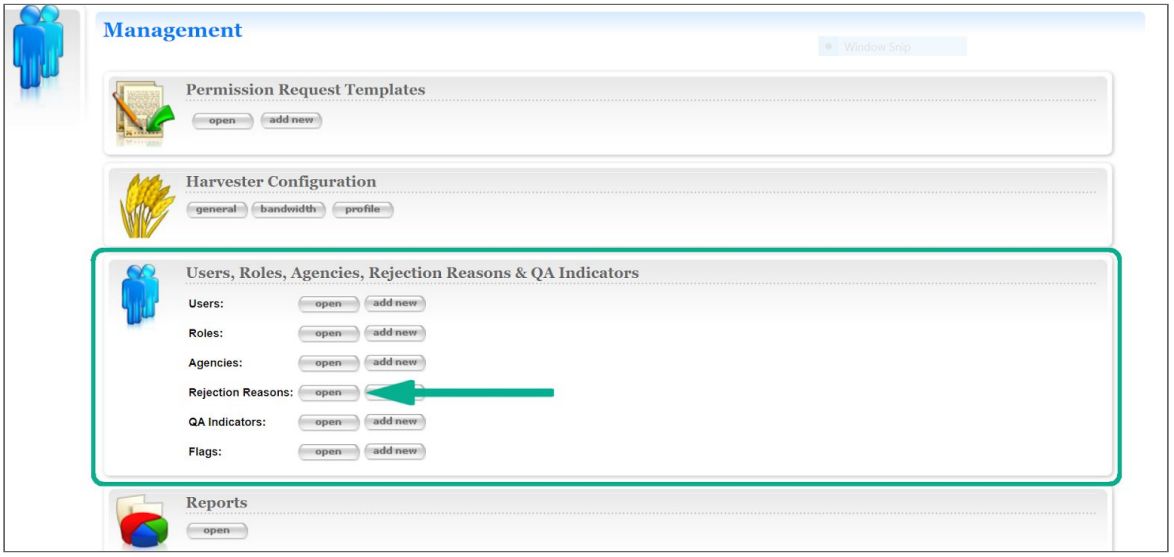
What you will learn [description, profiles, scheduling, crawl monitoring, quality review]

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

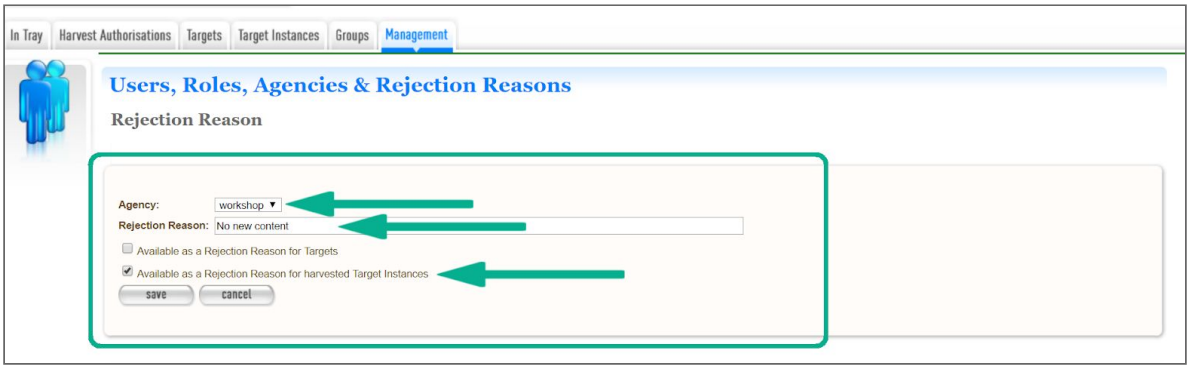
Setup Rejection Reasons

A rejection reasons can be assigned to Targets and Target Instances to state why it has been rejected. For example, a Target might be rejected for curatorial reasons, or a harvested Target Instance might be rejected for technical reasons.
For more information on Rejection Reasons in WCT, see the documentation.

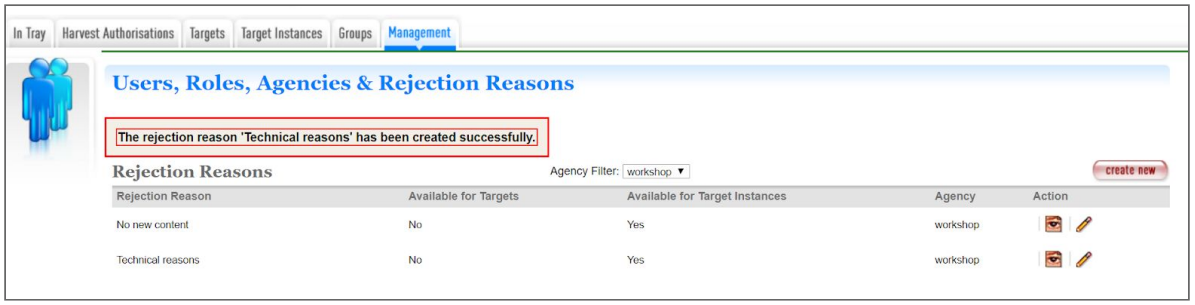
1. Navigate to the Management tab in WCT.
2. Click the Open option next to Rejection Reasons.



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



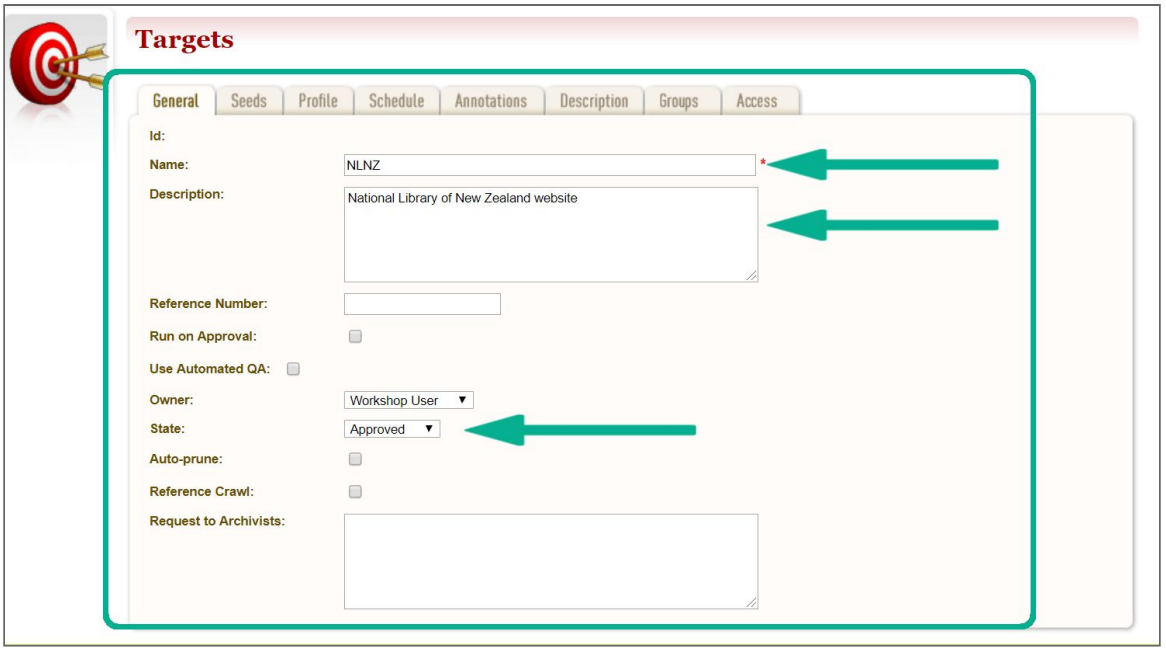
4. Create another Rejection Reason, and enter the following detail, and click Save.
   Agency: workshop
   Rejection Reason: Technical reasons
   Available as a Rejection Reason for harvested Target Instances: <tick>



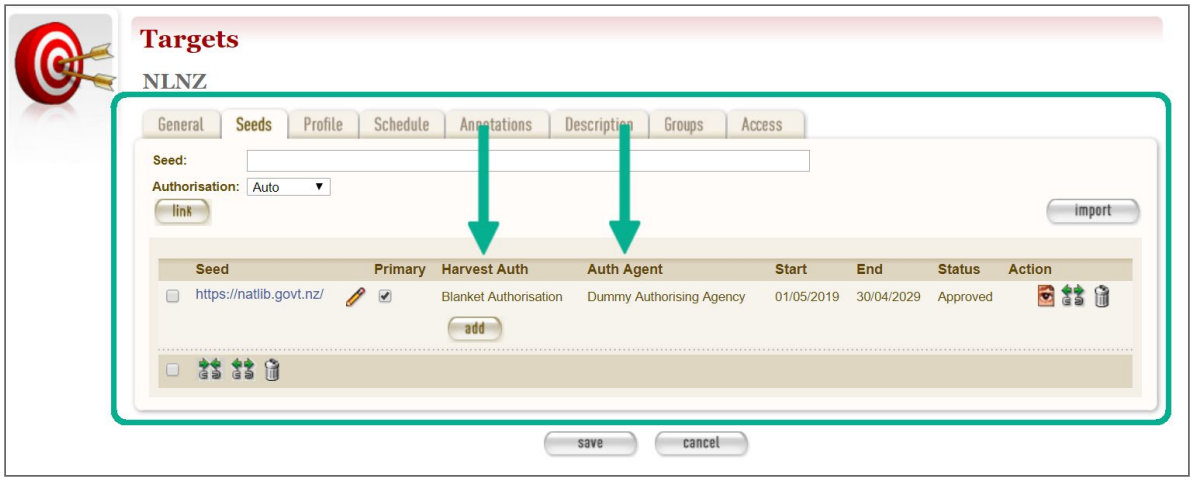
Setup Target

A Target is a record that describes the thing you intend to harvest from the Web. It contains metadata descriptions, seed URLs, how and when the thing will be crawled.
For more information on WCT Targets, see the documentation.

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

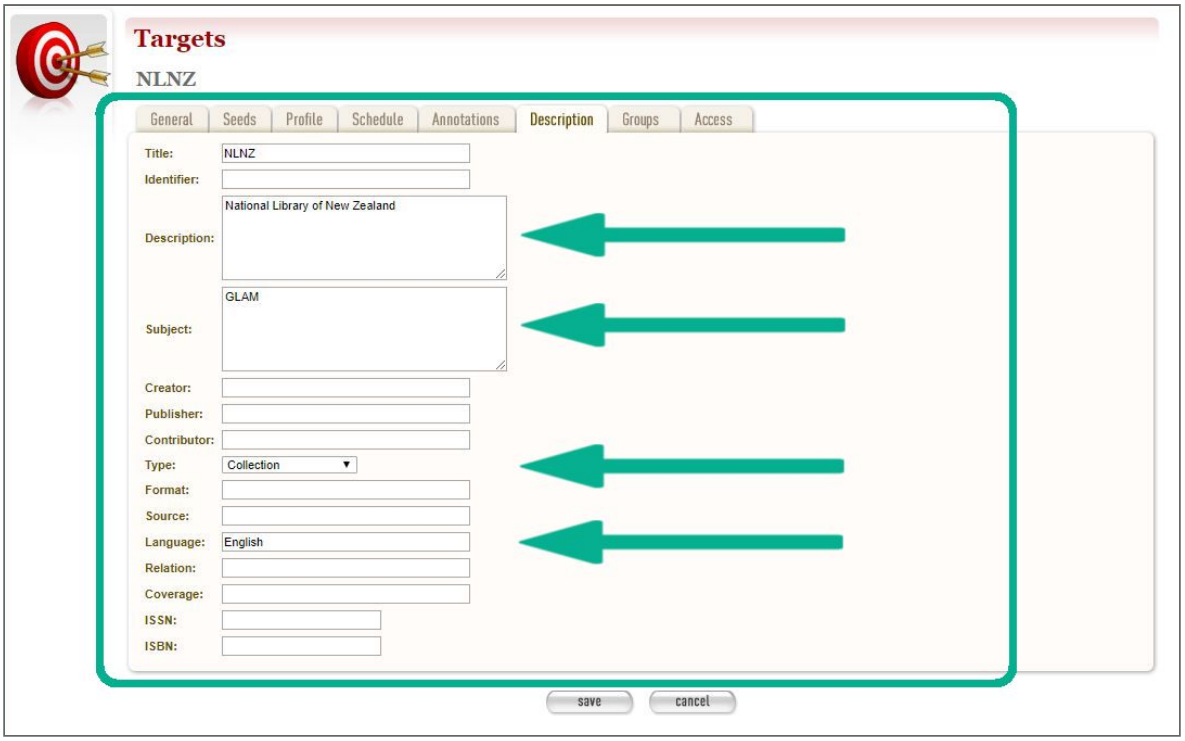


4. Navigate to the Seeds tab. If the Authorisation option is set to auto, then any new seeds will automatically be linked to Harvest Authorisations with matching URL patterns. Enter the following detail, then click the Link button.
   Seed: <your institutional website>
   Authorisation: Auto

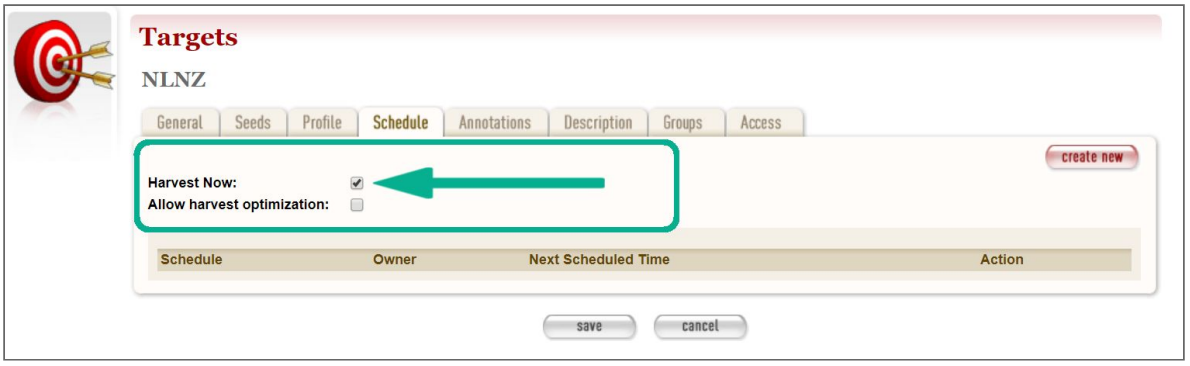


5. Navigate to the Profile tab, and ensure the following detail is set.
   Harvester Type: HERITRIX3
   Base Profile: Default - workshop

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>



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

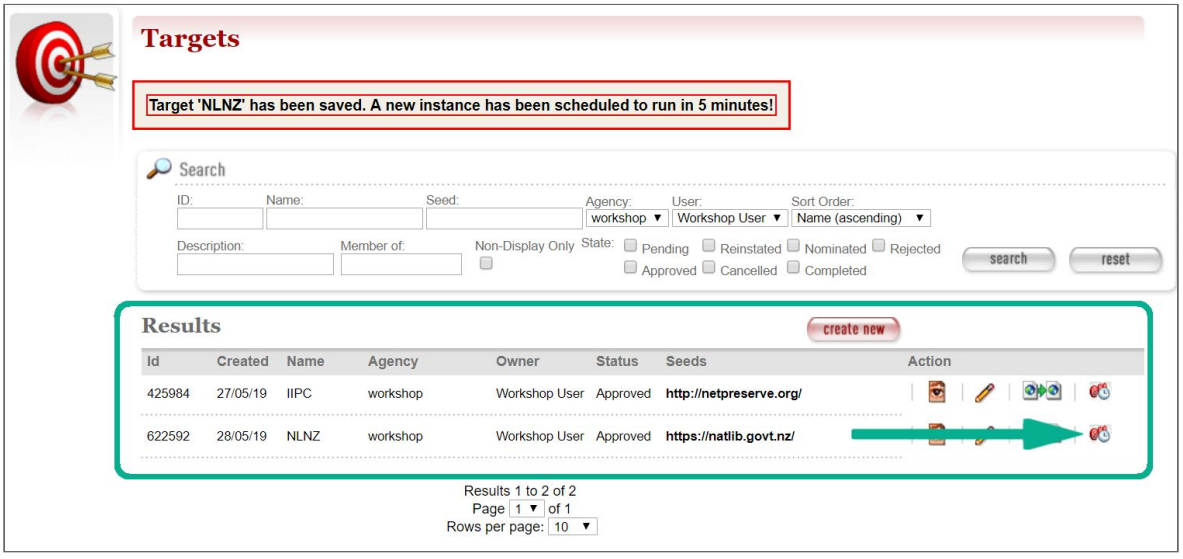


8. Save the Target.

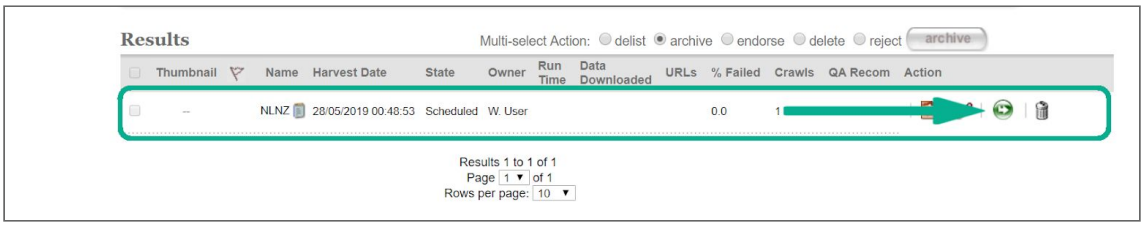
Monitor Target Instance

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.

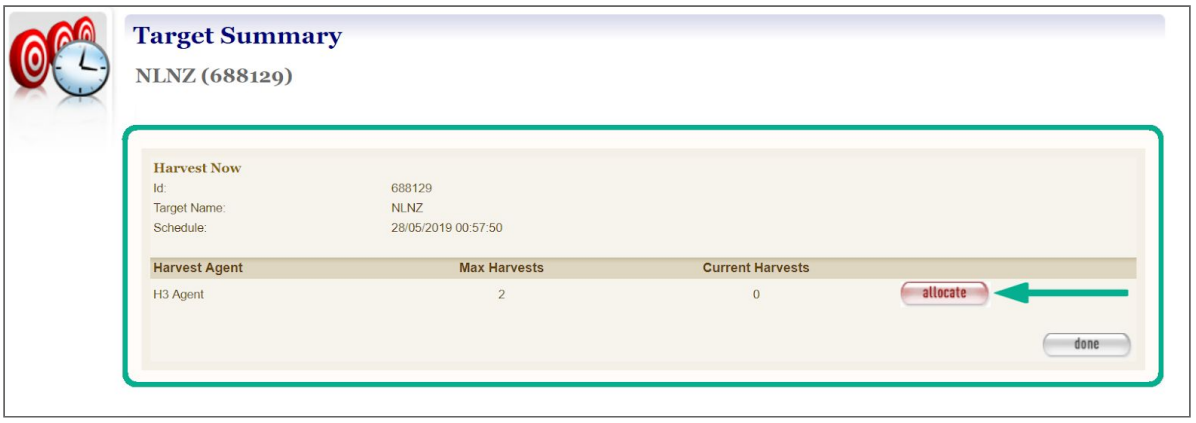
1. For the newly created Target, click the View Target Instances icon. This will search for all the Target Instances for this Target.



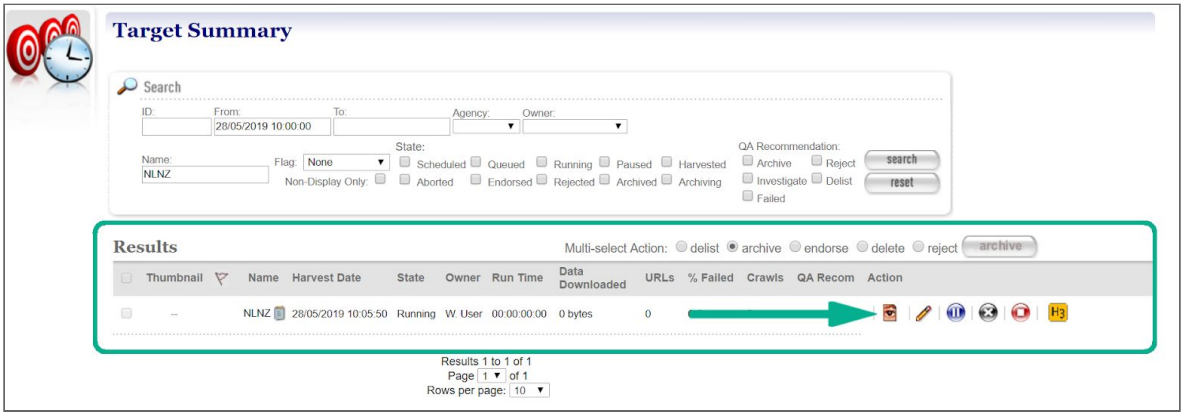
2. Observe the scheduled Target Instance, and click the Harvest Now icon.



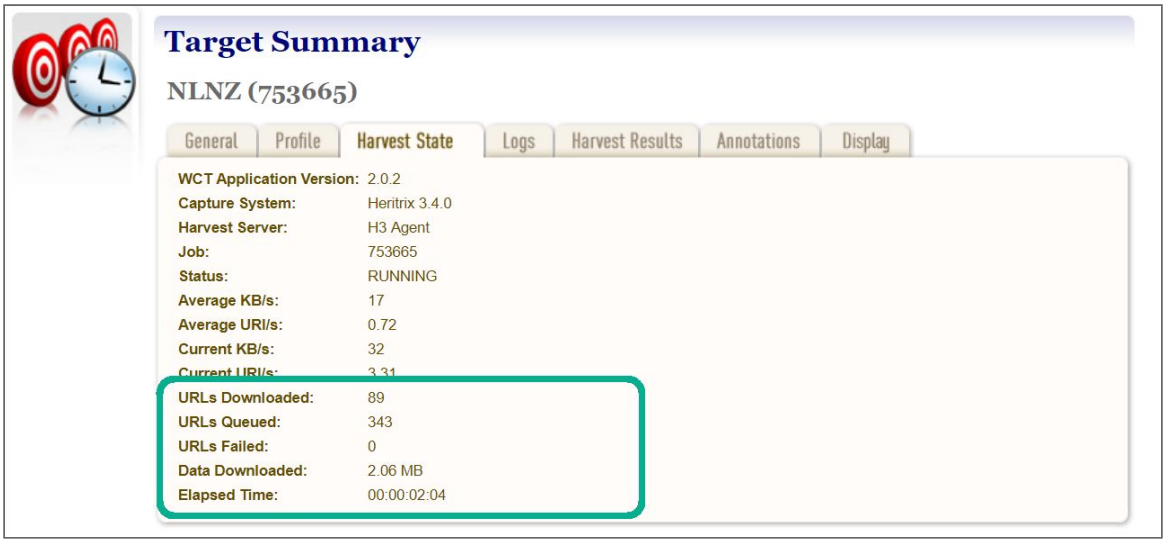
3. Assign the Target Instance to the available Harvest Agent, by clicking Allocate.



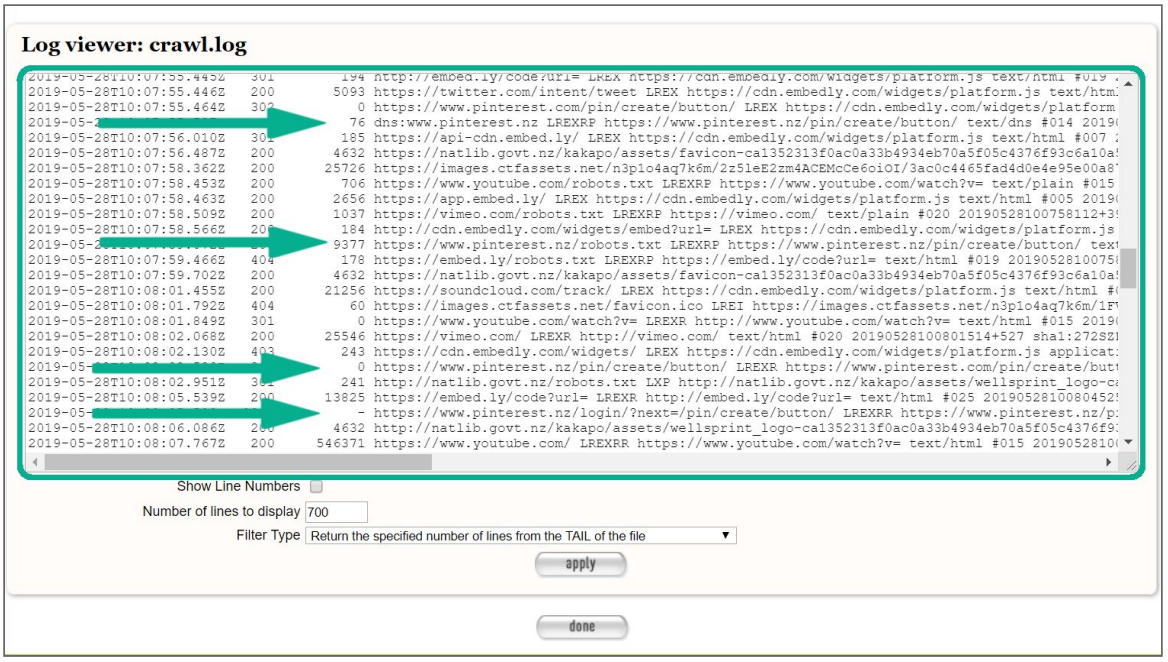
4. Observe the scheduled Target Instance move into a Running state. Click the View icon.



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*



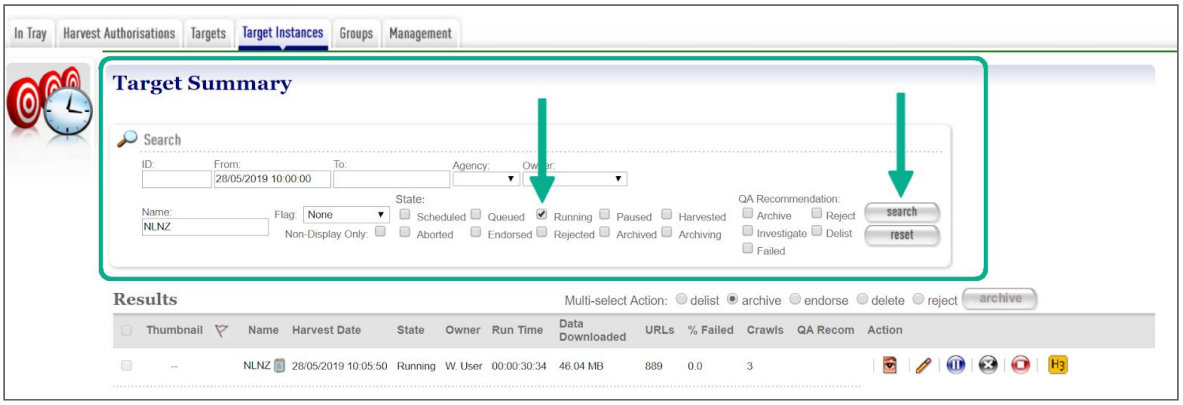
6. Navigate to the Logs tab, and click on the View link for crawl.log.
7. Scroll through the crawl log output. For this tutorial, locate any URLs from a hostname that you did not expect or want to be included in this crawl.
   For example, URLs from the Pinterest host.



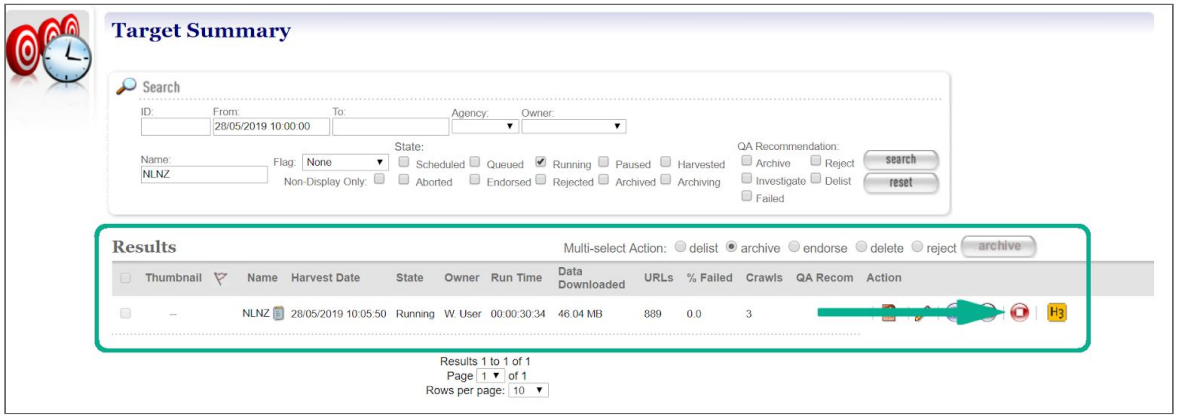
8. To filter the crawl log output, and only see entries for your unwanted hostname, set the Filter Type to 'Return all lines matching regex', enter the hostname surrounded by the any character pattern **.*** and click Apply.
   *For example, **.*pinterest.****

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 Instance tab, tick the Running state filter, and click Search.



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.

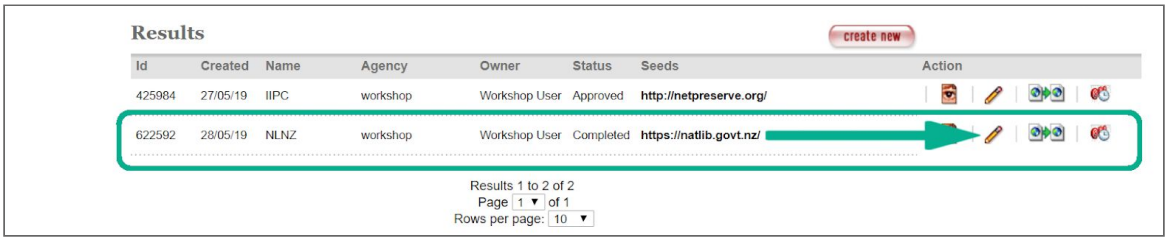


12. Stopping the Target Instance will keep the harvested content for review. Aborting the Target Instance will discard the harvested content and will not be available for review.

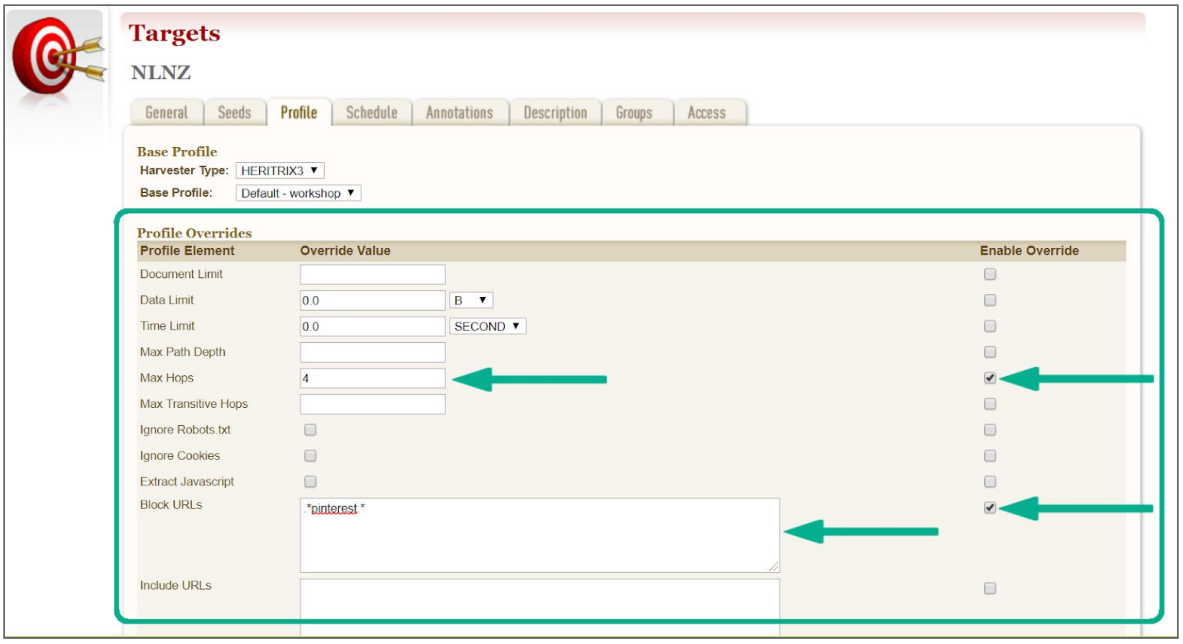
Adjust the Target

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.



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.**



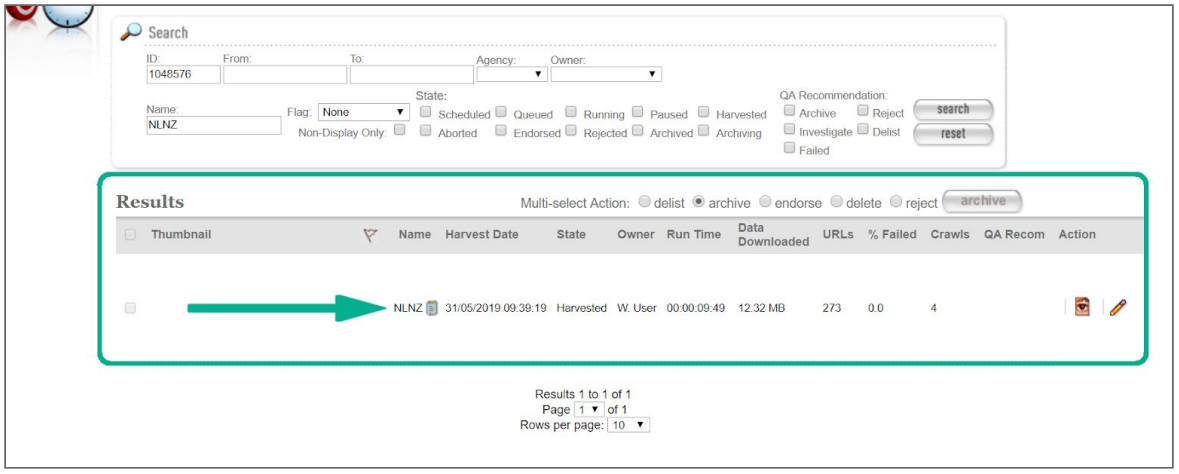
4. Navigate to the Schedule tab, tick the Harvest Now box, and click Save.
5. Click the View Target Instances icon next to the Target.
6. Start the Target Instance using the Harvest Now option.

Compare and review the crawl

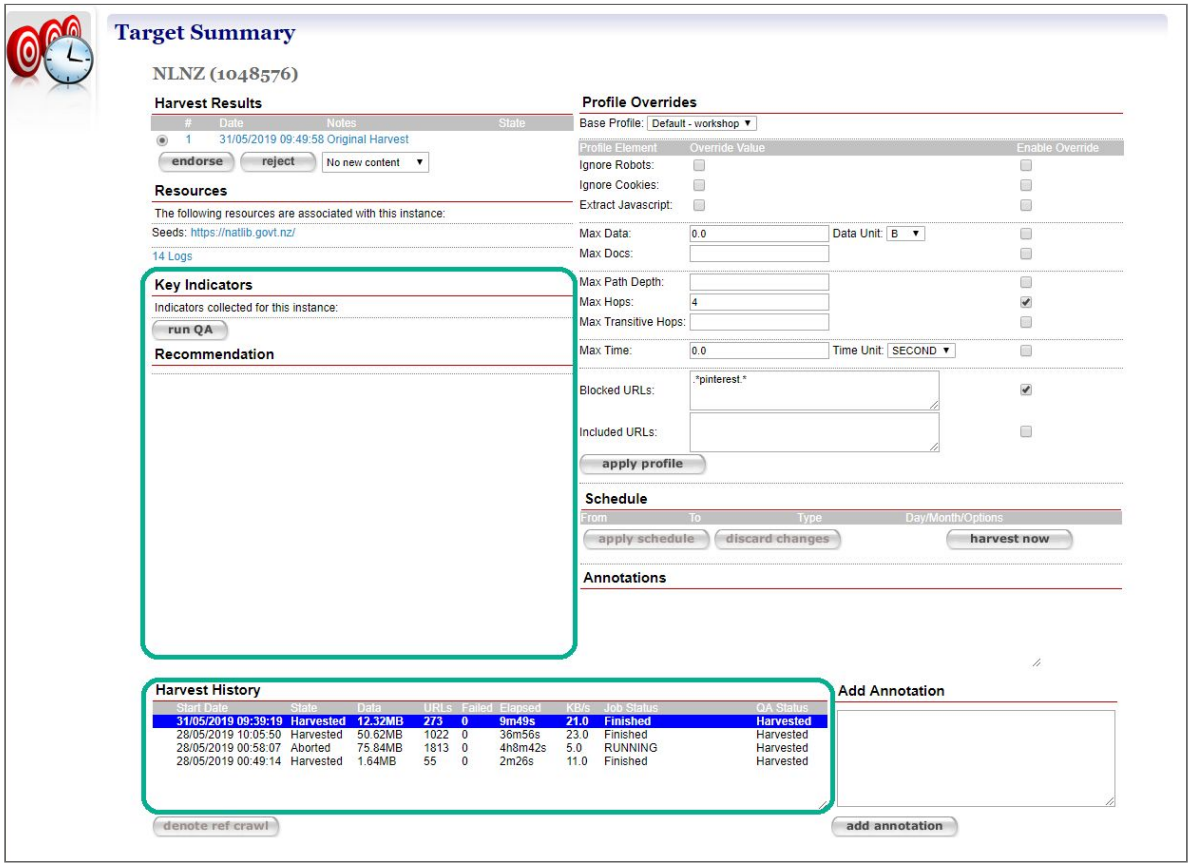
1. View the running Target Instance. Monitor the crawl log to ensure you don't see any URLs from your now excluded hostname. Filter on the regular expression
   *Regular Expression: .*pinterest.**

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.

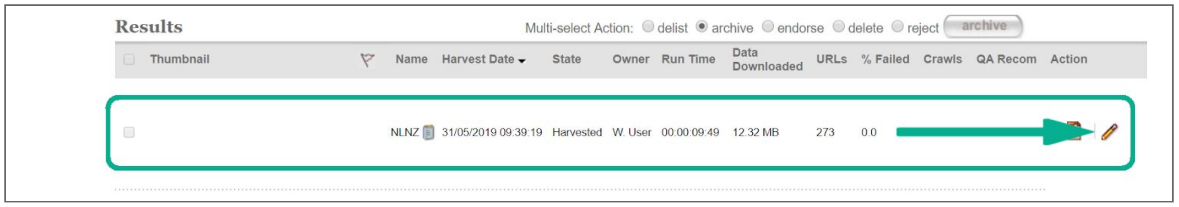
3. Once the Target Instance has moved into the *Harvested* state, click on the Target Instance name to go the the Target Summary screen.



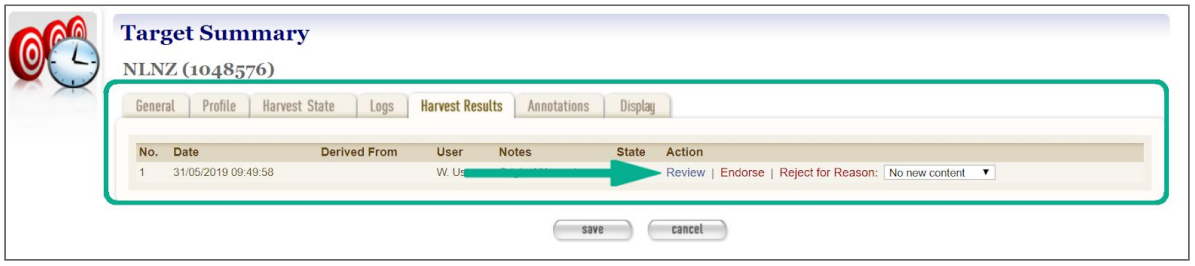
4. The Target Summary has several areas to assist with quality review. The Harvest History table shows comparable statistics for the previous crawls. The Key Indicators and Recommendation sections are blank because the indicators must be setup in the Management tab.
For more information on the Target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



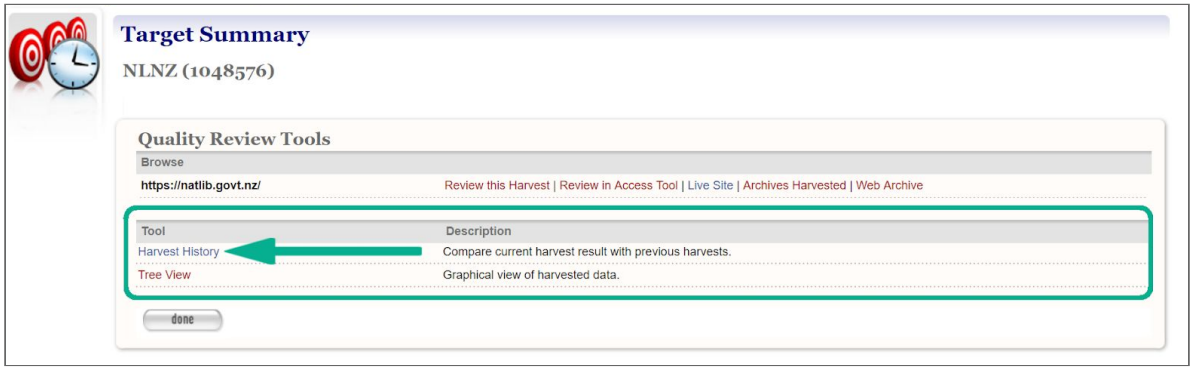
5. Navigate back to the Target Instance tab, and click the Edit icon for the same Target Instance.



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



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



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.

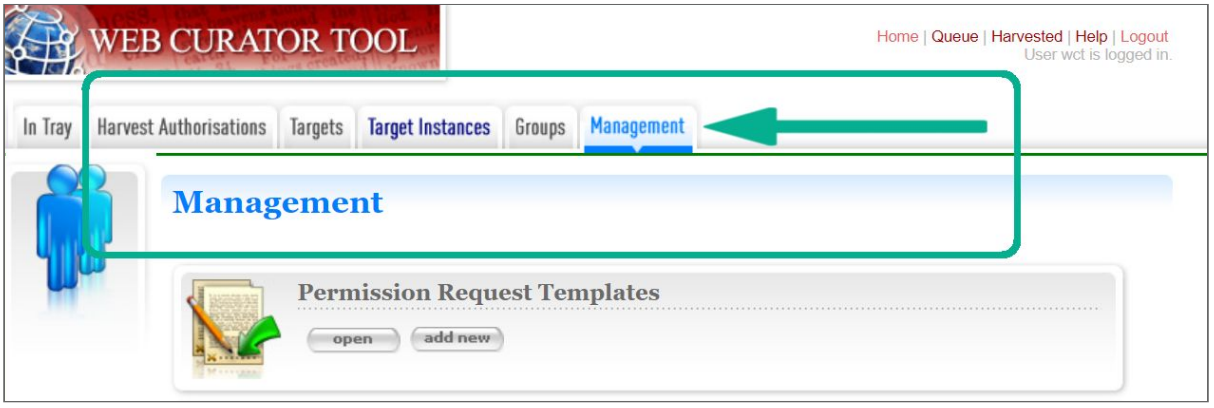
In many countries across the world deposit legislation allows libraries and other heritage institutions to harvest and store all websites belonging to their national domain. Of course what does and does not belong to the national domain will often be subject to debate, but in terms of authorisation the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

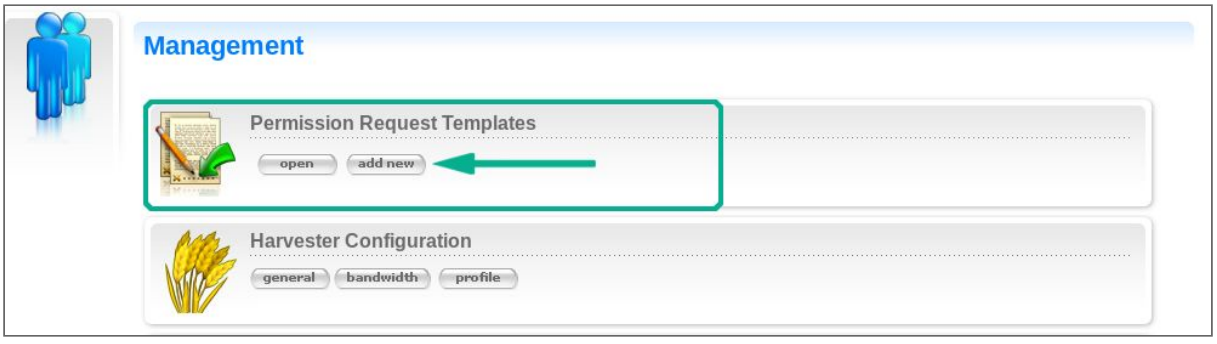
Create a Permission Request Template

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.

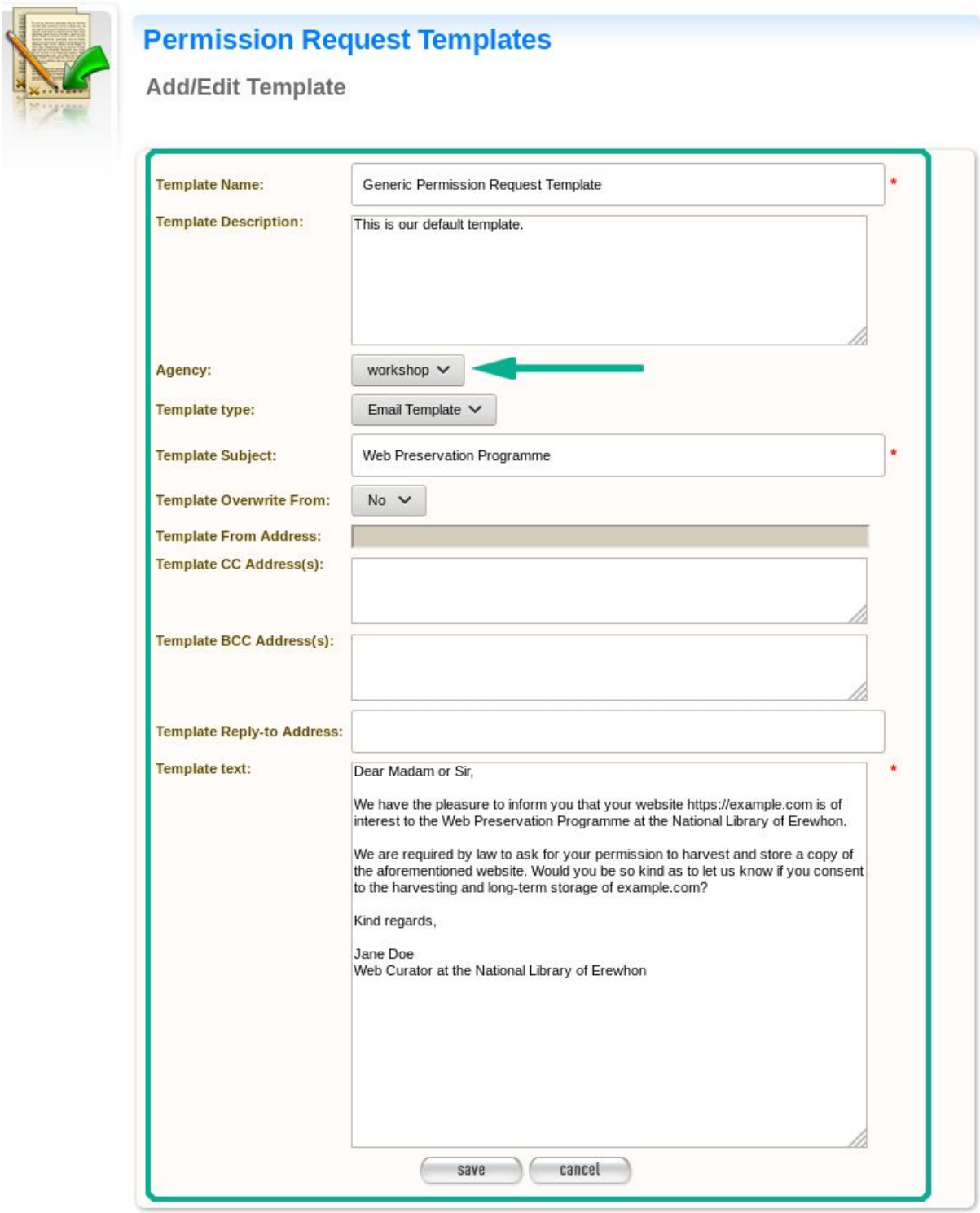
1. Navigate to the Management tab.



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>

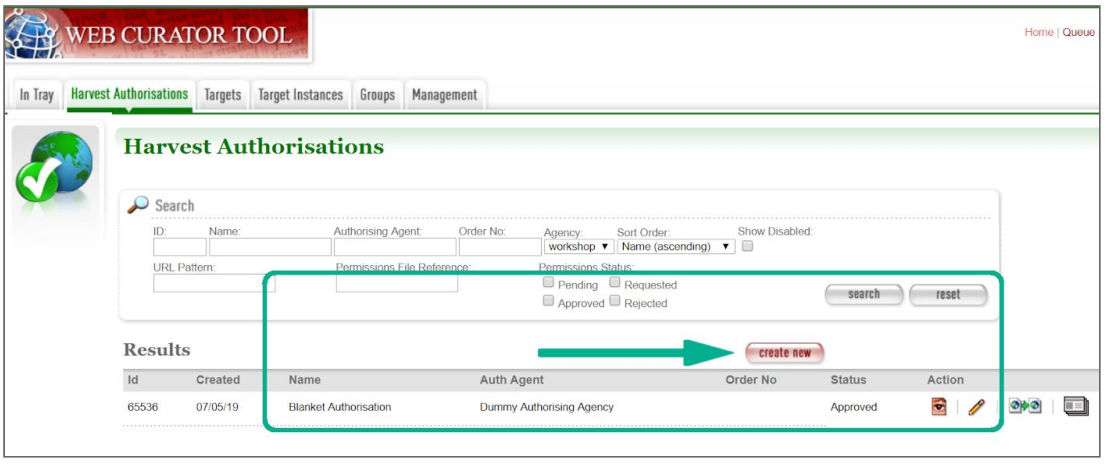


Create Harvest Authorisation

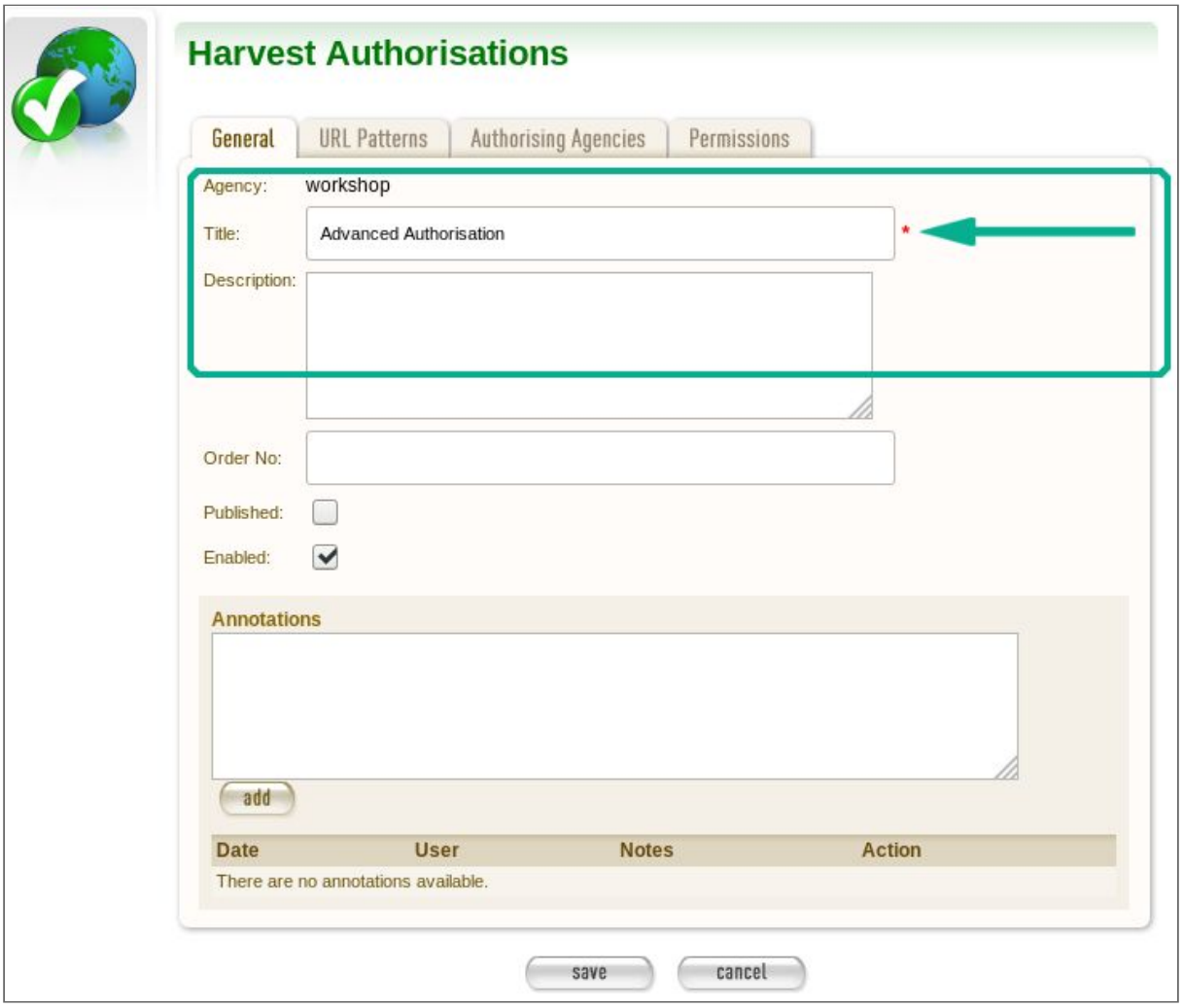
Next, we're going to create a new Harvest Authorisation.

1. Navigate to the Harvest Authorisations tab.

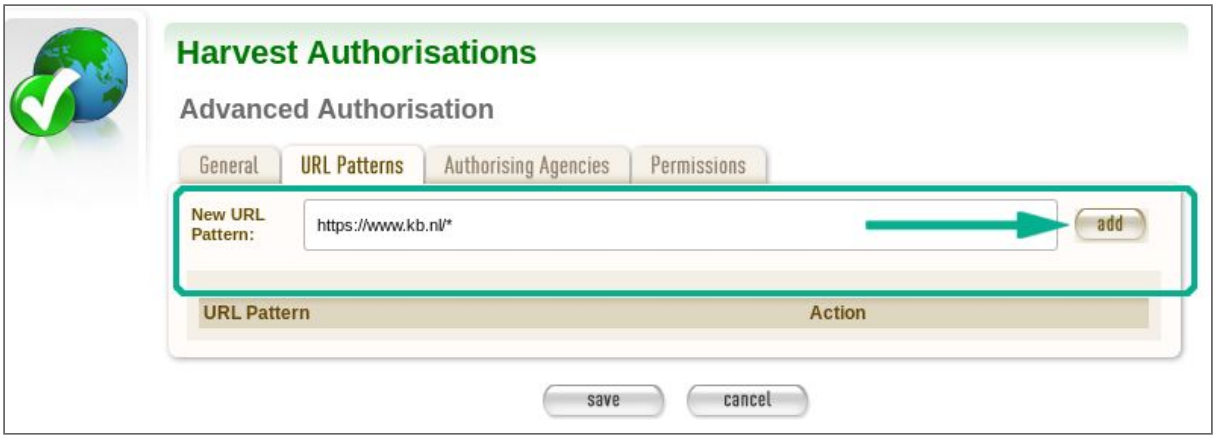
2. Create a new Harvest Authorisation.



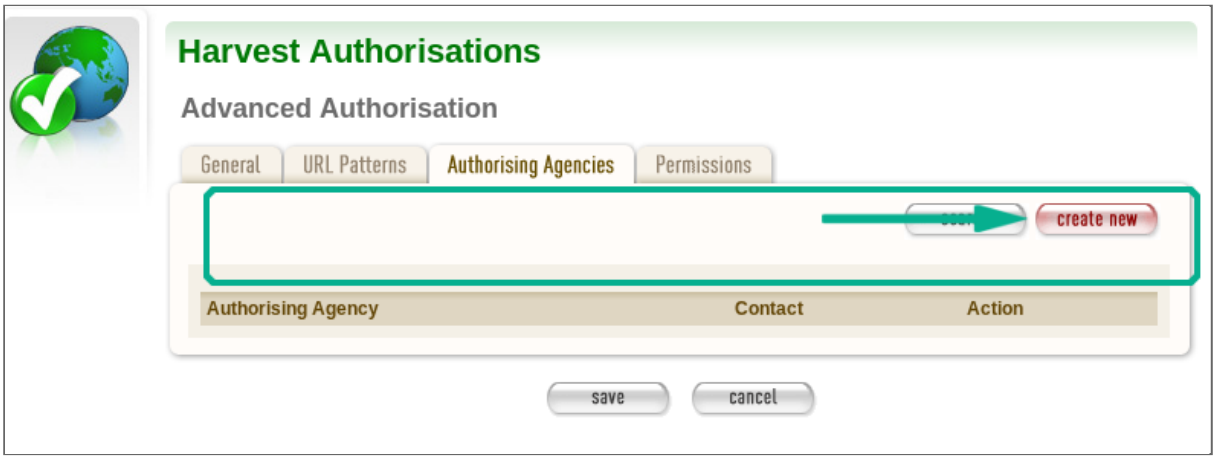
3. Under the General tab enter:
Title: Advanced Authorisation



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/.



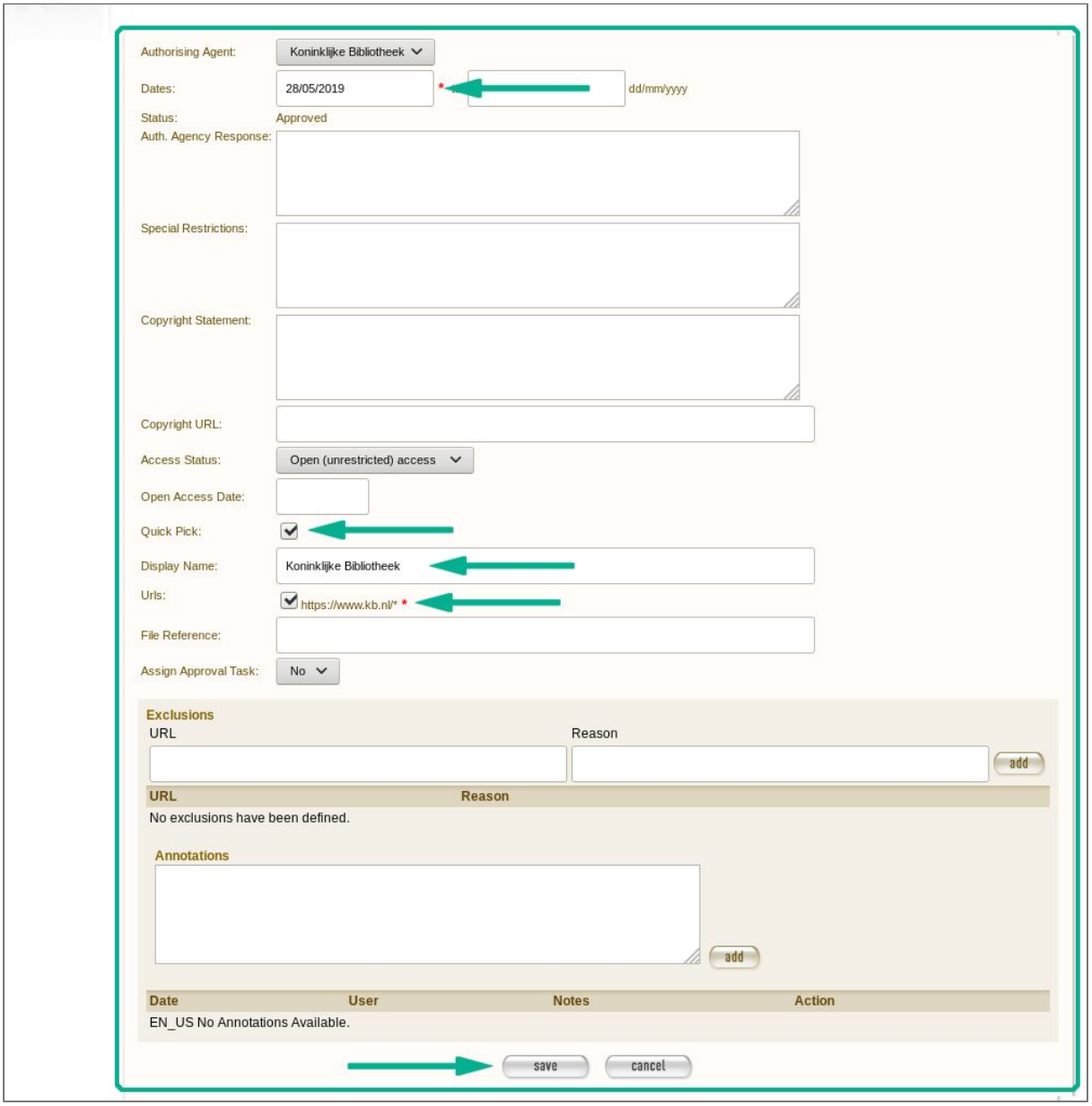
5. Create a new Authorising Agency under the next tab.



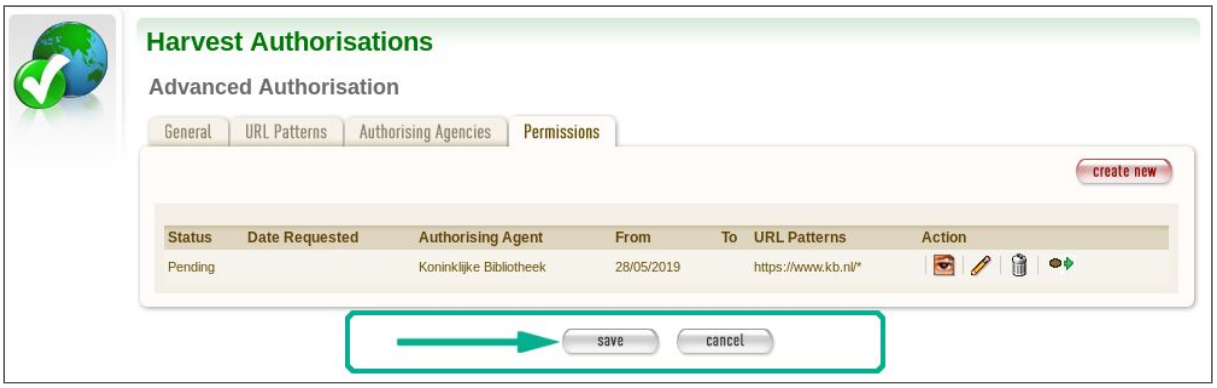
6. Fill in the contact details of the site owner. For the purposes of this workshop, use the name of your institution. The email address doesn't have to be real.

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

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>



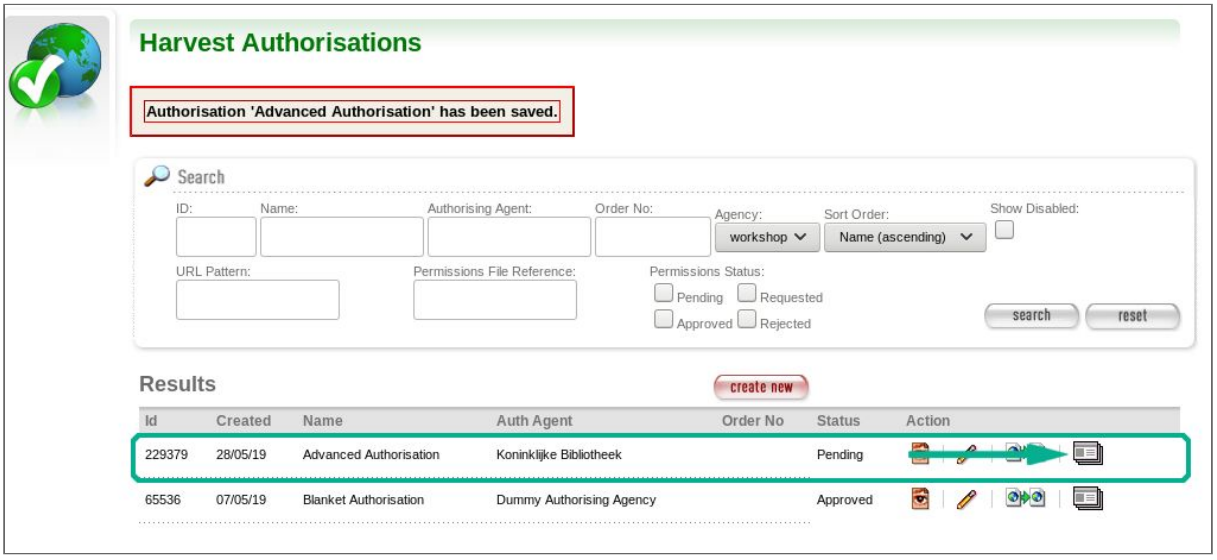
9. Finally, save the Harvest Authorisation.



Send Permission Request

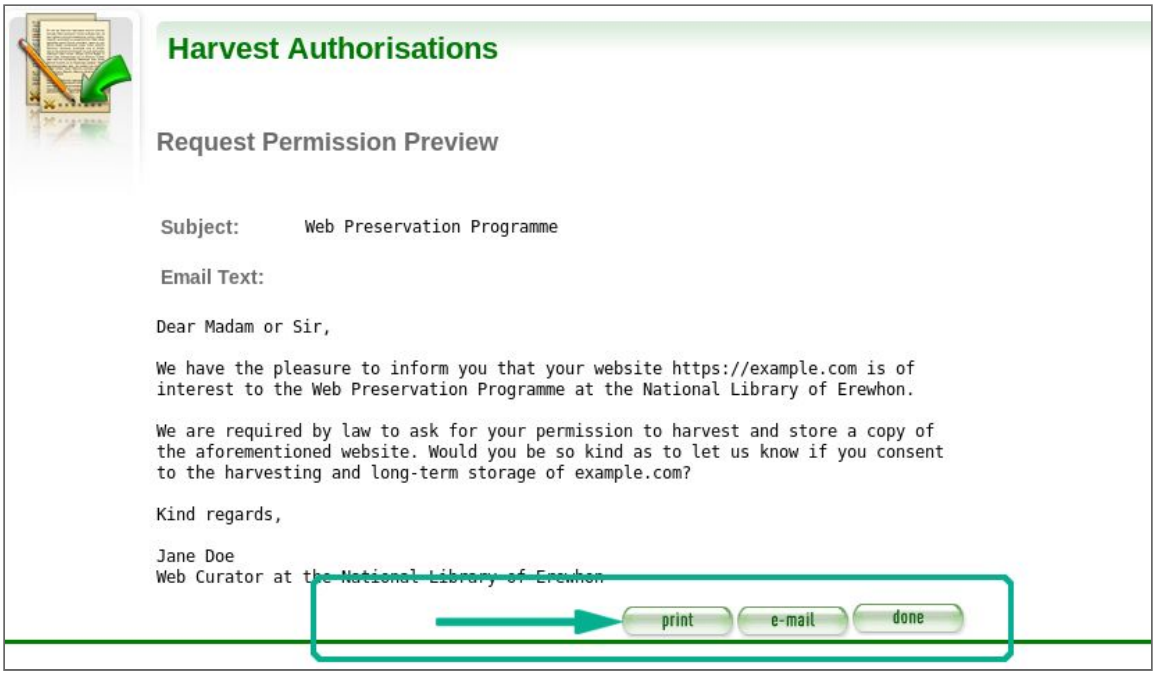
We are now ready to send the permission request to the website owner.

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

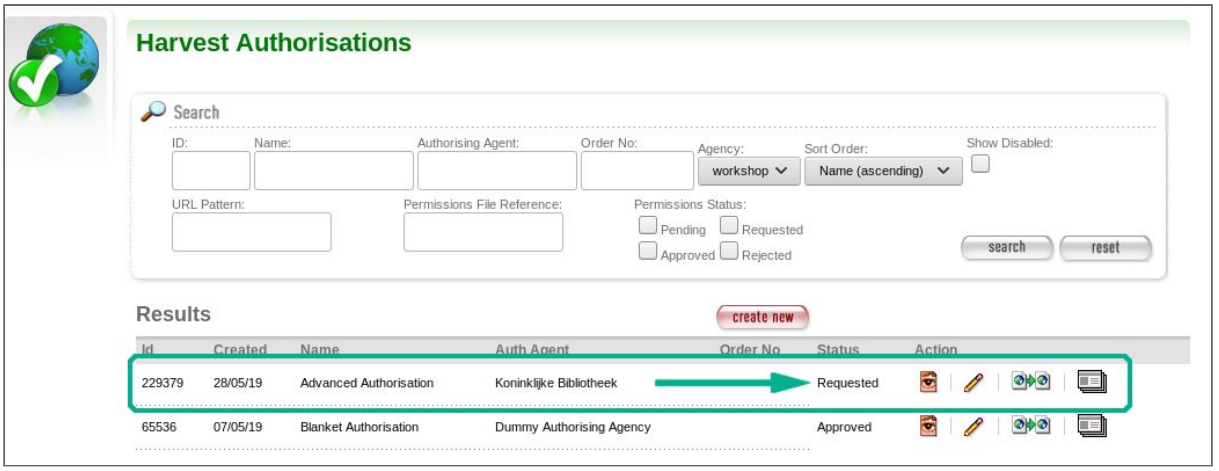


2. You are now on a screen where you can choose which request to send to the site owner. In our case there's only one permission request template. Click on the letter icon.

3. 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.



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



Revisiting the Harvest Authorisation

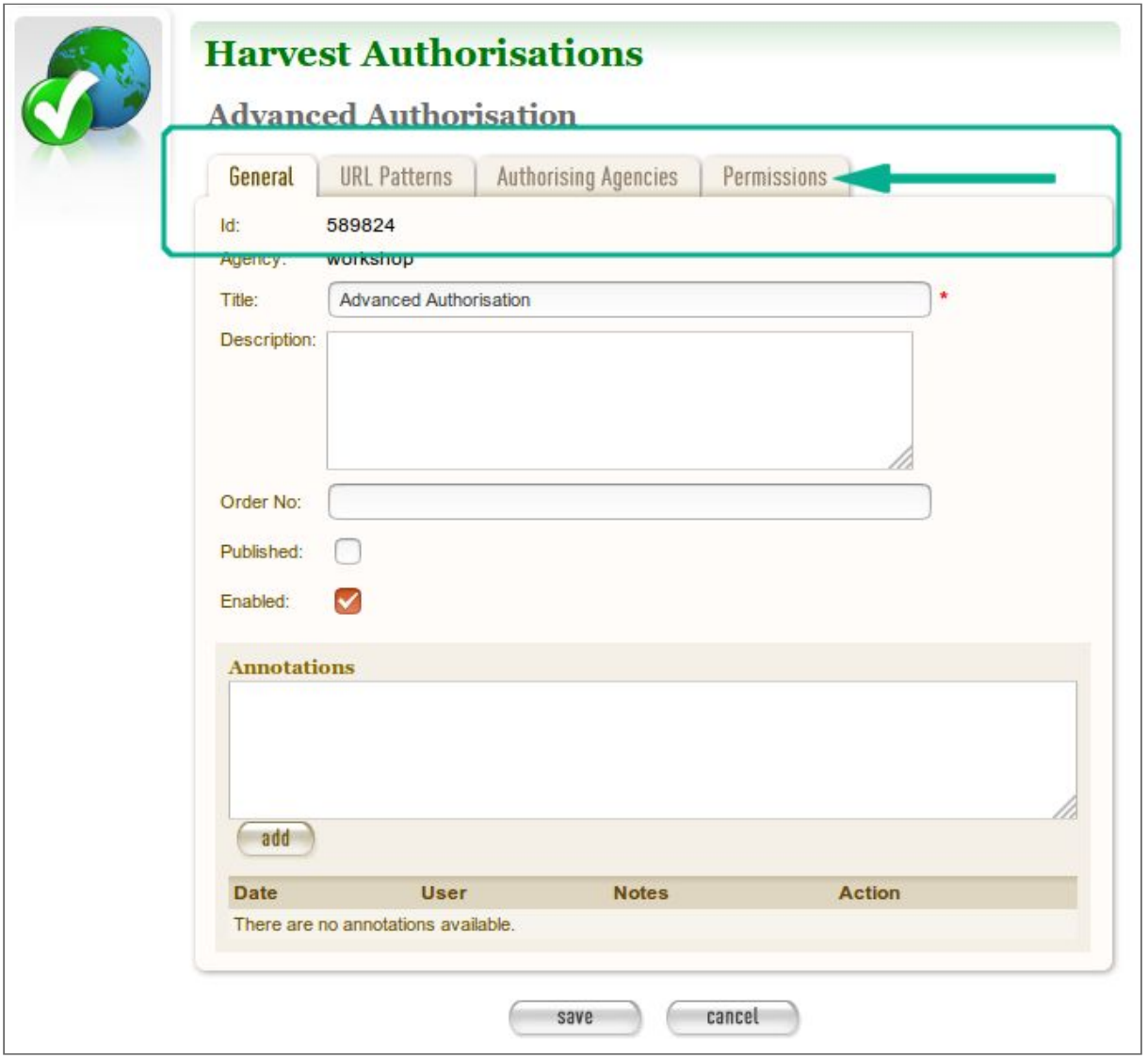
After some time the site owner will hopefully respond positively to your permission request. When that happens you will need to revisit the Harvest Authorisation to record the permission and change the status of the Permissions to *Approved.*

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.

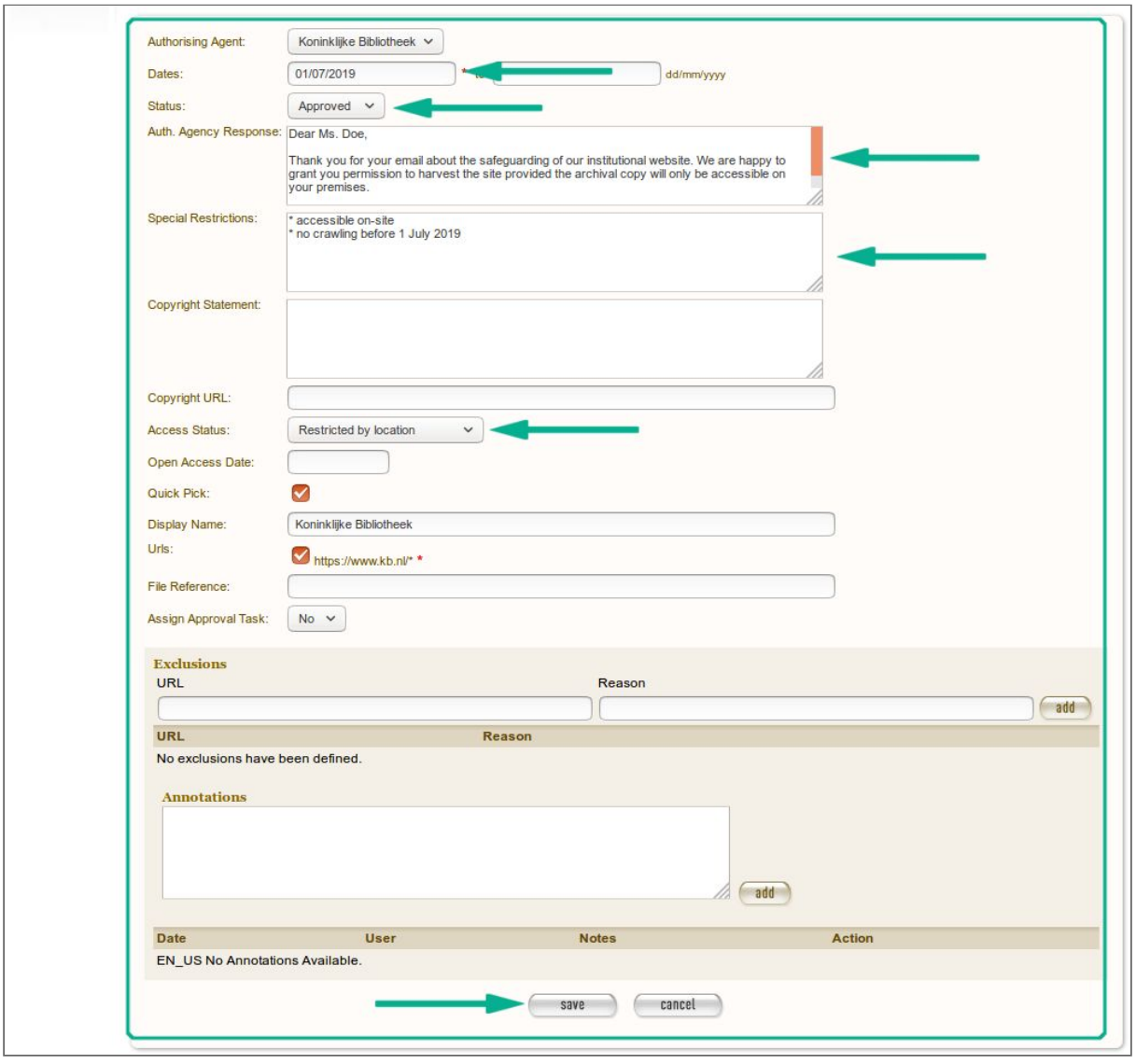
1. Navigate to the Harvest Authorisations tab.
2. Open your Advanced Authorisation in edit mode.

3. Go to the Permissions tab.

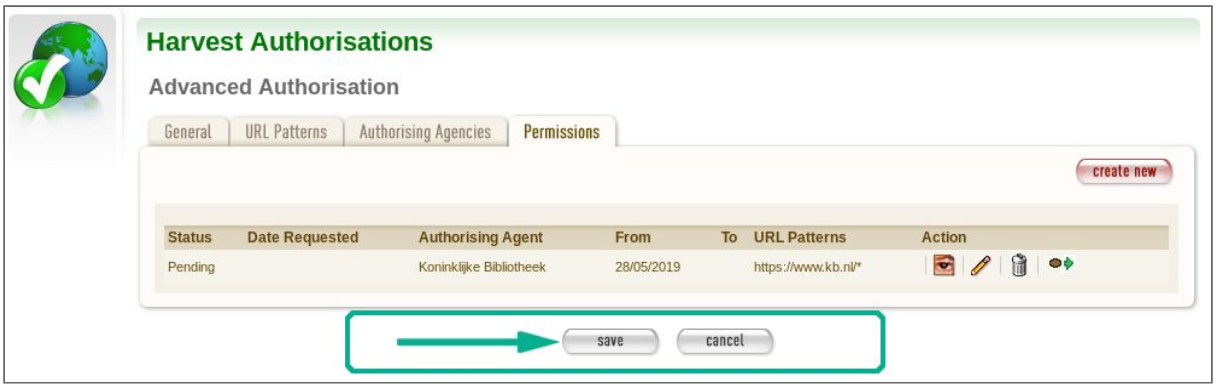


4. Open the existing permission in edit mode.

5. Enter the following detail, and click Save.
   Dates: <a future date> to <blank>
   Status: Approved
   Auth. Agency Response: <paste the body of the email you've received from the site owner>
   Special Restrictions: <note the restrictions on crawl start time and access>
   Access Status: Restricted by location



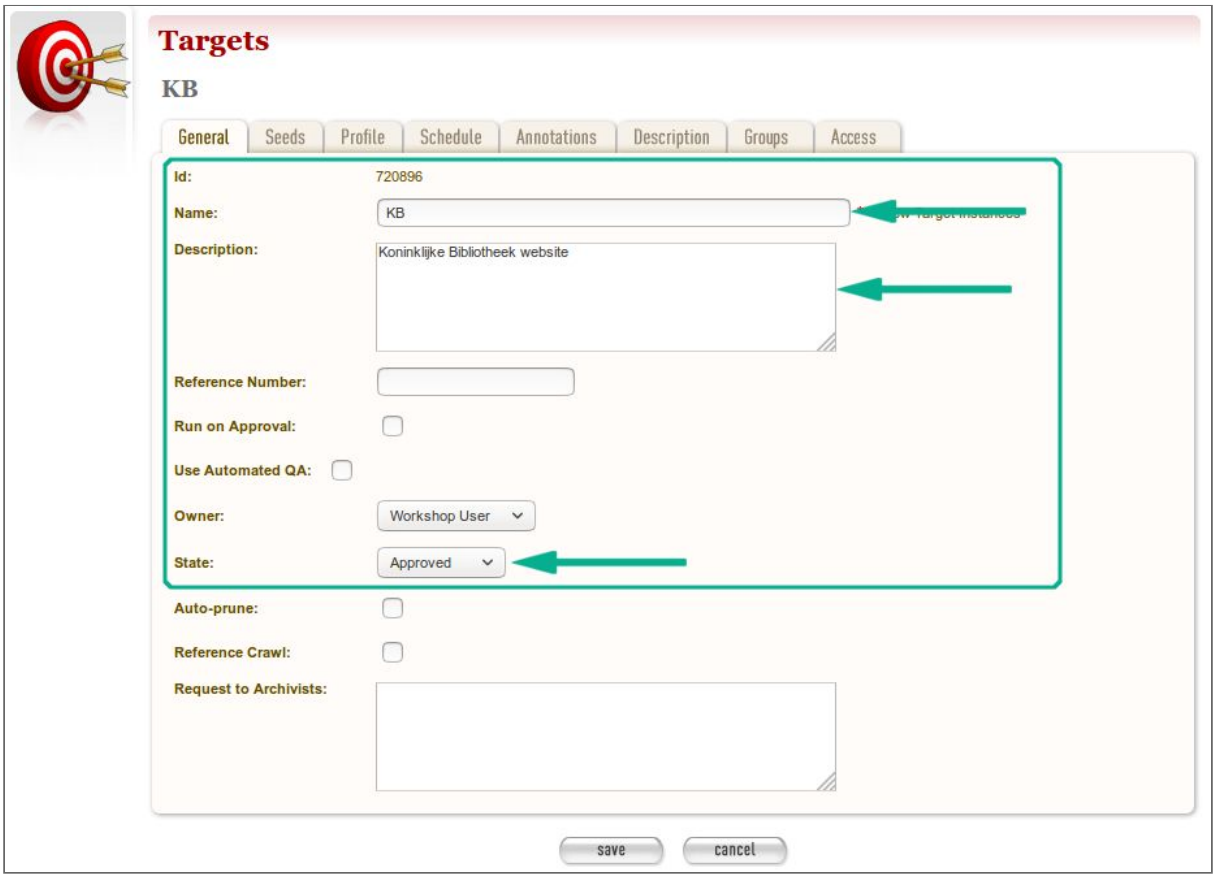
6. Finally, save the Harvest Authorisation.



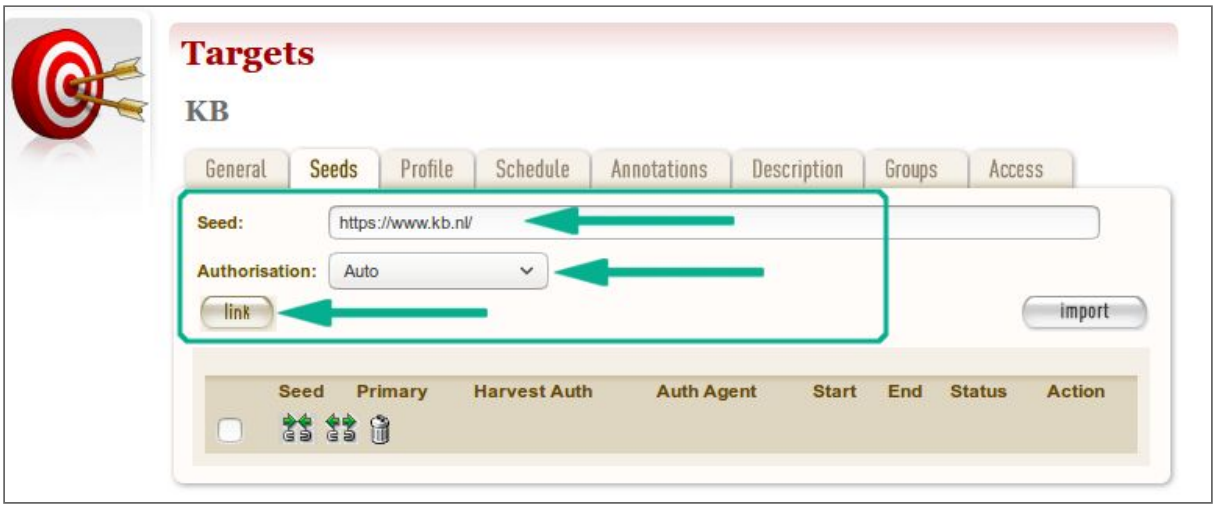
Attempting a crawl

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

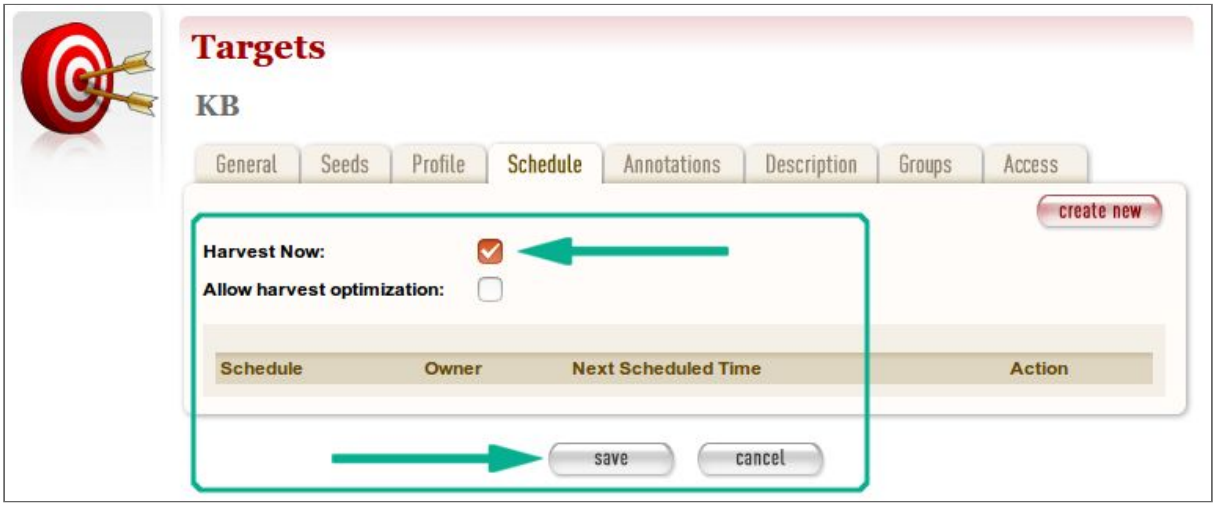
1. Navigate to the Targets tab.
2. Create a new Target.
3. In the General tab enter the following detail
   Name: <short name of your institution>
   Description: <full name of your institution> website
   State: Approved



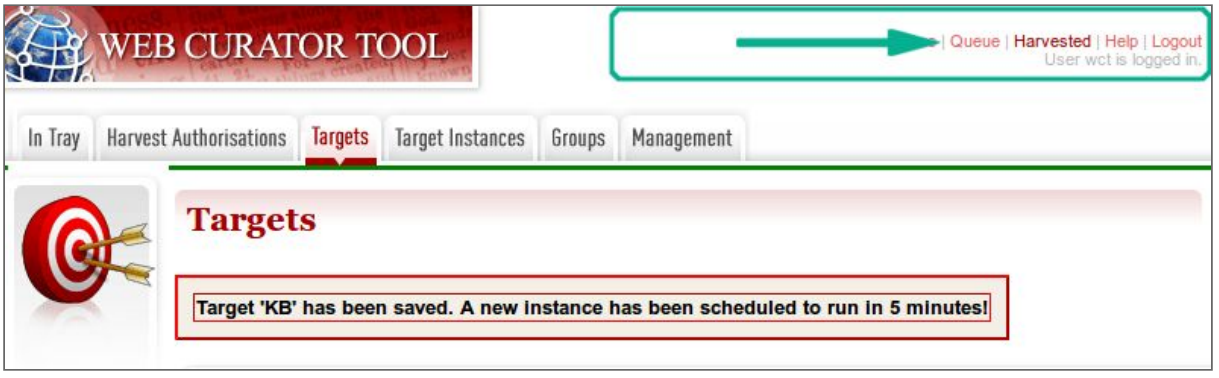
4. In the Seeds tab, add a URL that conforms to the URL pattern that you added under the Permissions tab of the Harvest Authorisation earlier.
5. From the Authorisation dropdown, select the one that corresponds to the name you set as Display Name under the Permissions tab earlier (it's listed here, because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



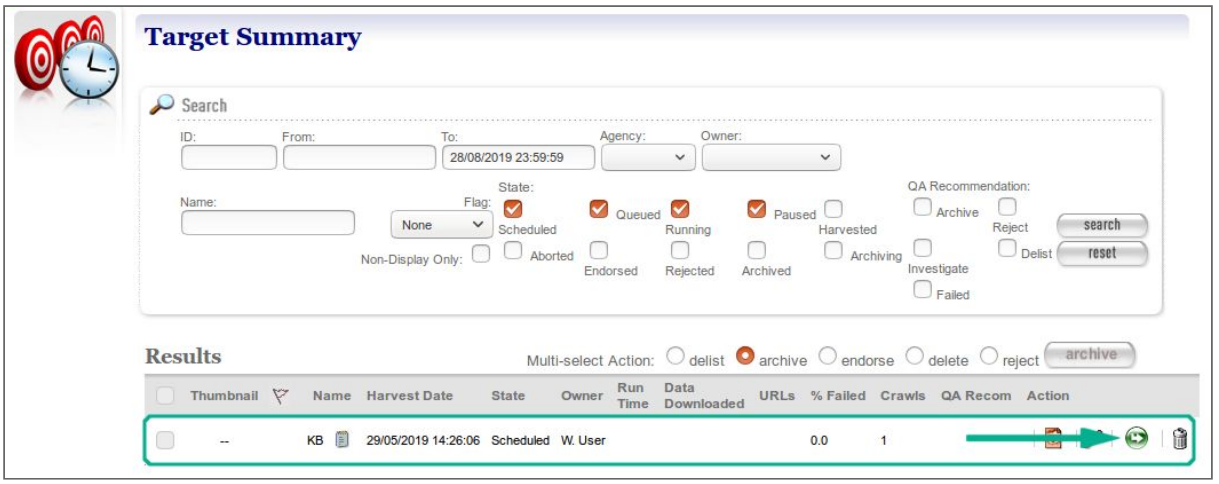
7. In the Schedule tab, select Harvest Now and save the Target.



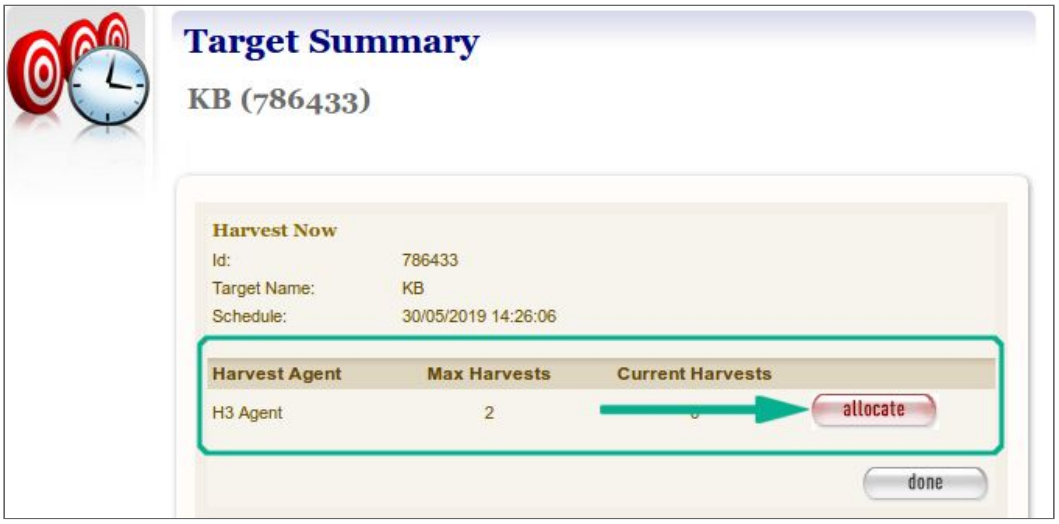
8. A crawl will be started in 5 minutes. Of course we are a bit impatient, so we want to start it now. To do that go to the Queue.



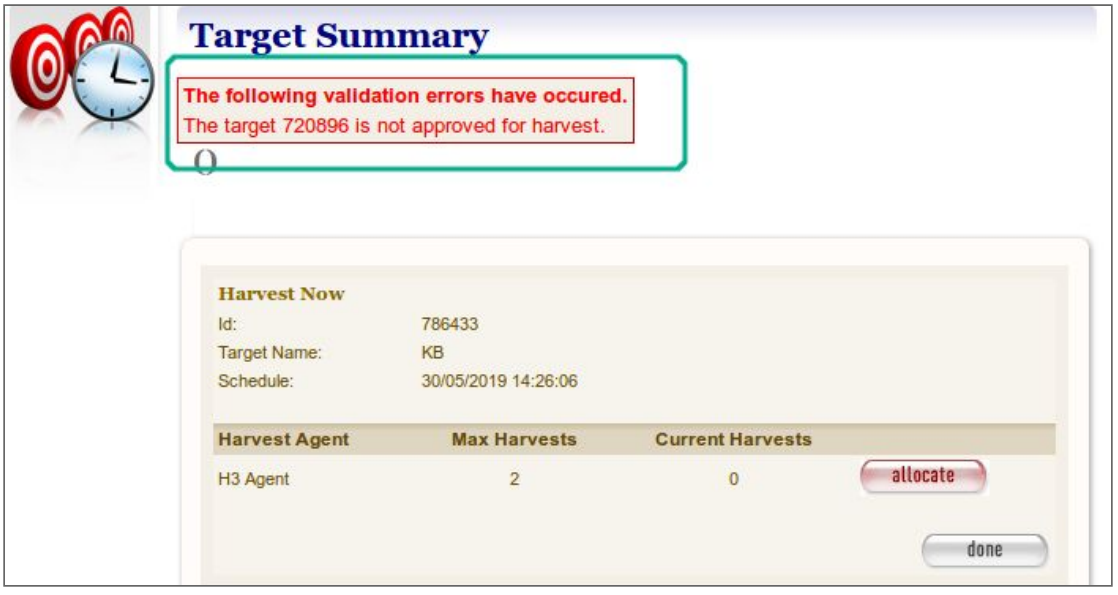
9. And click on the Harvest Now button of our newly created Target Instance.



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



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.



Tutorial D - Advanced Crawling and Quality Review

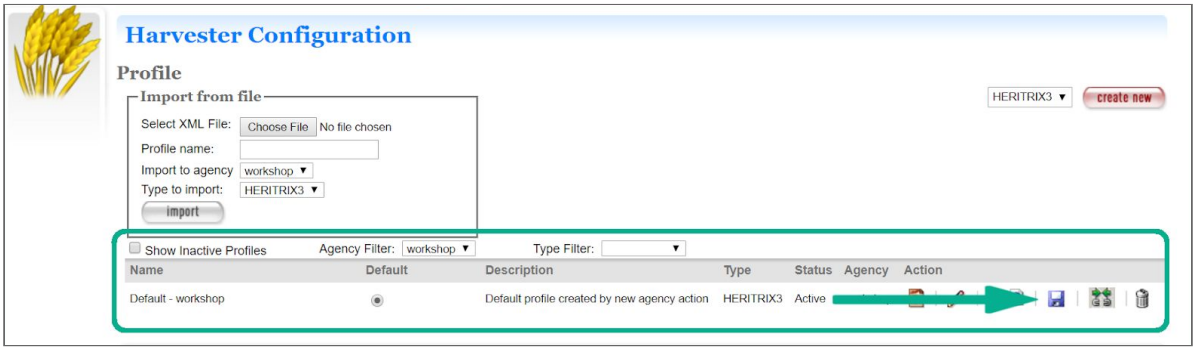
What you will learn [seeds, profiles, scheduling, groups, crawl monitoring, Heritrix 3 scripting, quality review]

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.

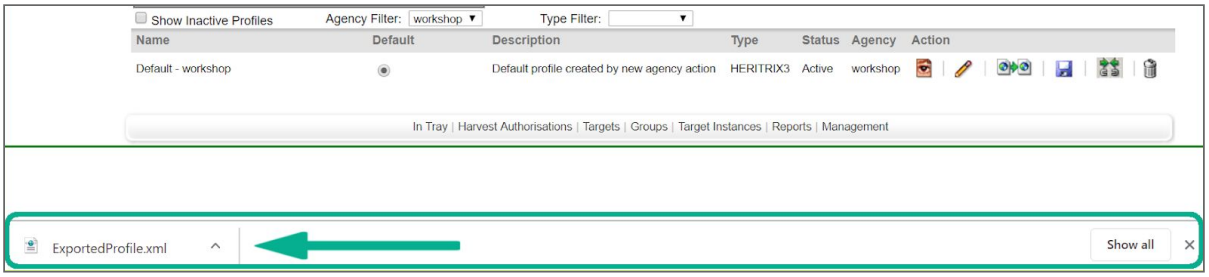
Importing Profiles

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.

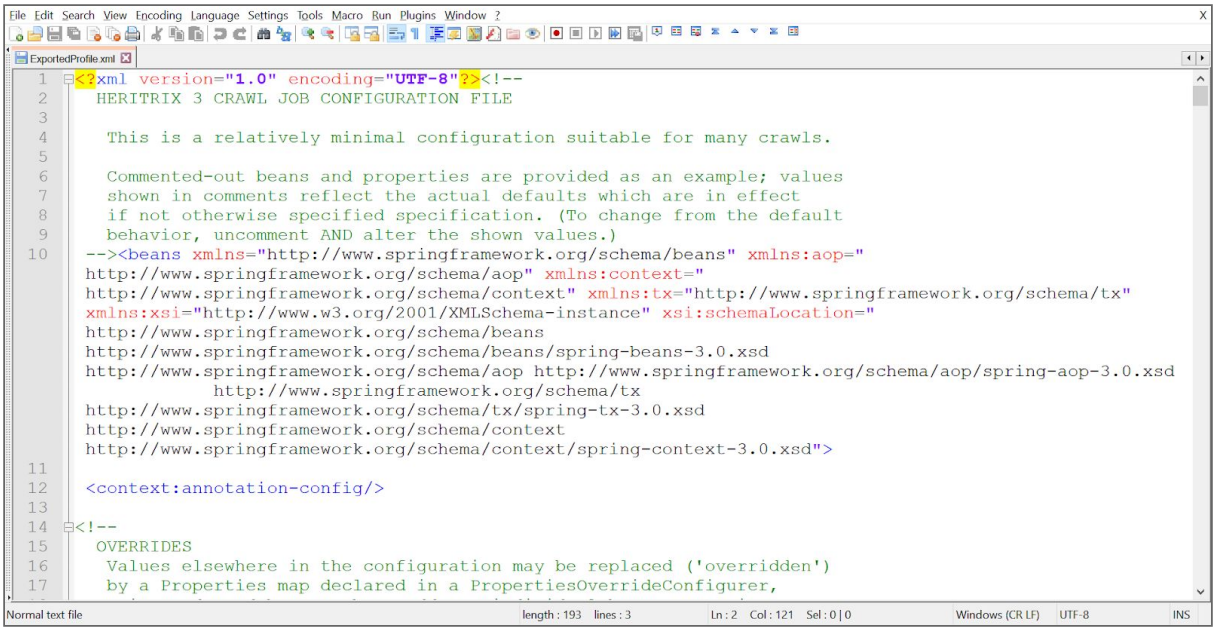
1. Navigate to the Management tab in WCT
2. Click the Profile option under Harvester Configuration.
3. You should see the default profile listed - “Default - workshop”. Click the Expor ticon next to the default profile.



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



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



6. Choose one or more of the following examples and update the profile xml. Or make your own configuration change if you have knowledge of working with Heritrix profiles.
    a. Increase the maximum repetitions allowed in URL paths. Remove the xml comment syntax (<!--, -->), and increase the value to 10.

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

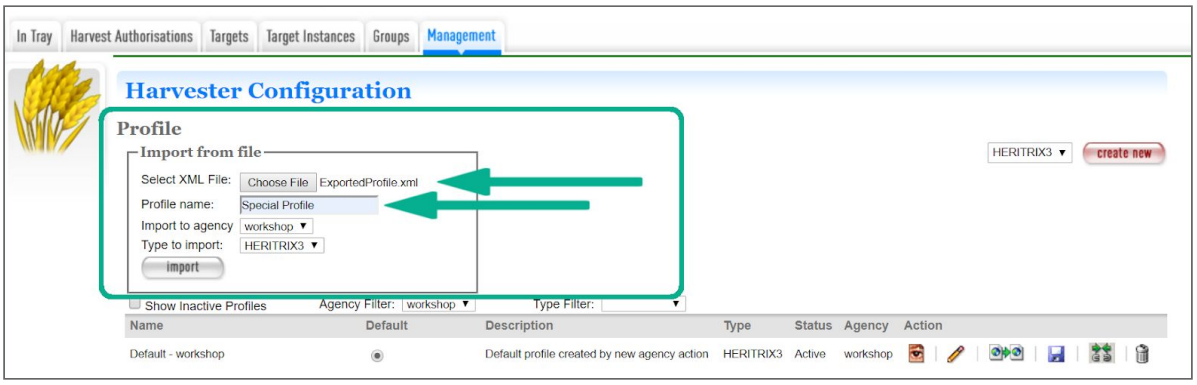
    b. Reduce the amount of time Heritrix waits before giving up on a failing URL. Remove the xml comment syntax (<!--, -->), and reduce the *retryDelaySeconds* value to 600 and the *maxRetries* value to 10.

```
<bean class="org.archive.crawler.frontier.BdbFrontier" id="frontier">

    <!-- <property name="snoozeLongMs" value="300000" /> -->
    <!-- <property name="retryDelaySeconds" value="900" /> -->
    <!-- <property name="maxRetries" value="30" /> -->
```

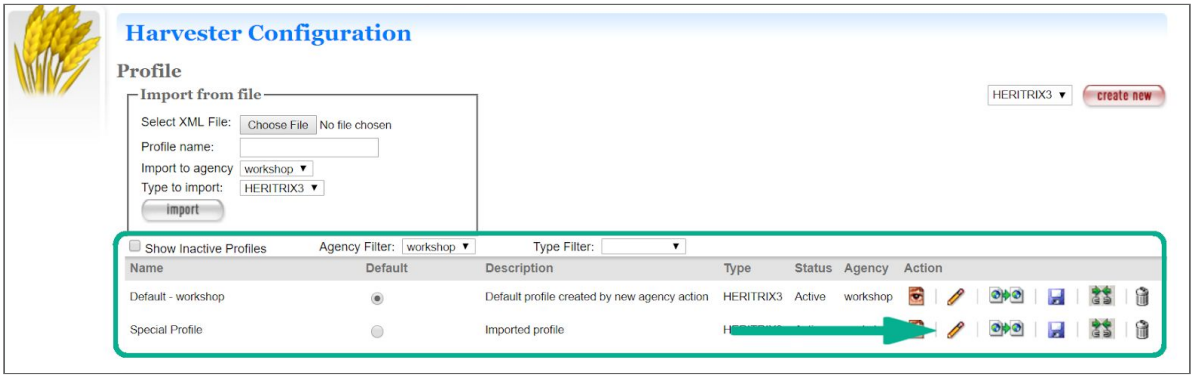
    c. Configure a http proxy for Heritrix to crawl through, and its credentials. Remove the xml comment syntax (<!--, -->), and enter the appropriate proxy configuration.

```
<bean class="org.archive.modules.fetcher.FetchHTTP" id="fetchHttp">

    <!-- <property name="httpProxyHost" value="" /> -->
    <!-- <property name="httpProxyPort" value="0" /> -->
    <!-- <property name="httpProxyUser" value="" /> -->
    <!-- <property name="httpProxyPassword" value="" /> -->
```

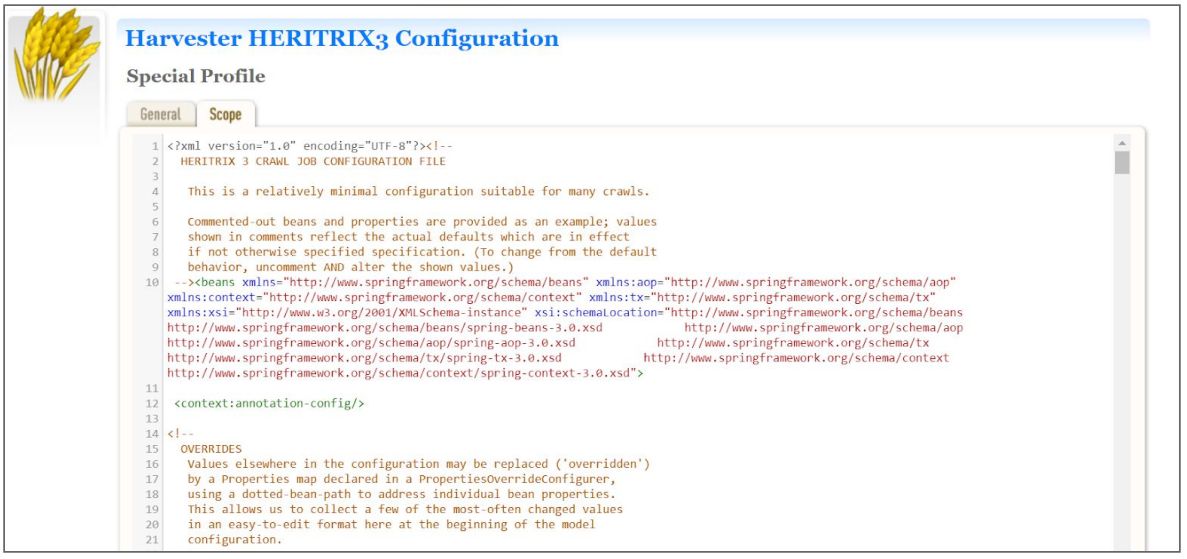
7. Once you have edited and saved the profile xml, import it back into WCT, using the following details.
   Select XML File: <browse and select the profile you just edited>
   Profile name: Special profile
   Import to agency: workshop
   Type to import: HERITRIX3



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



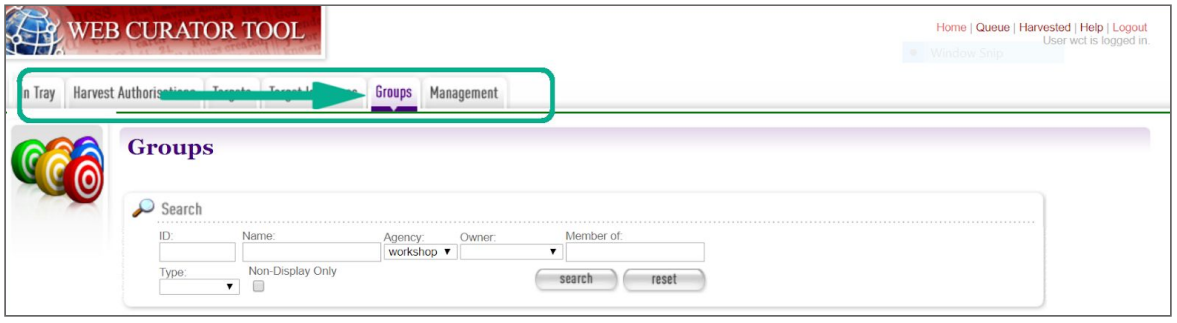
9. Navigate to the Scope tab.
   When editing/viewing an imported profile you will see an XML editor instead of the standard WCT Profile UI. This is done to remove the risk of WCT misreading a profile edited by a user, as the validity of the xml cannot be guaranteed.



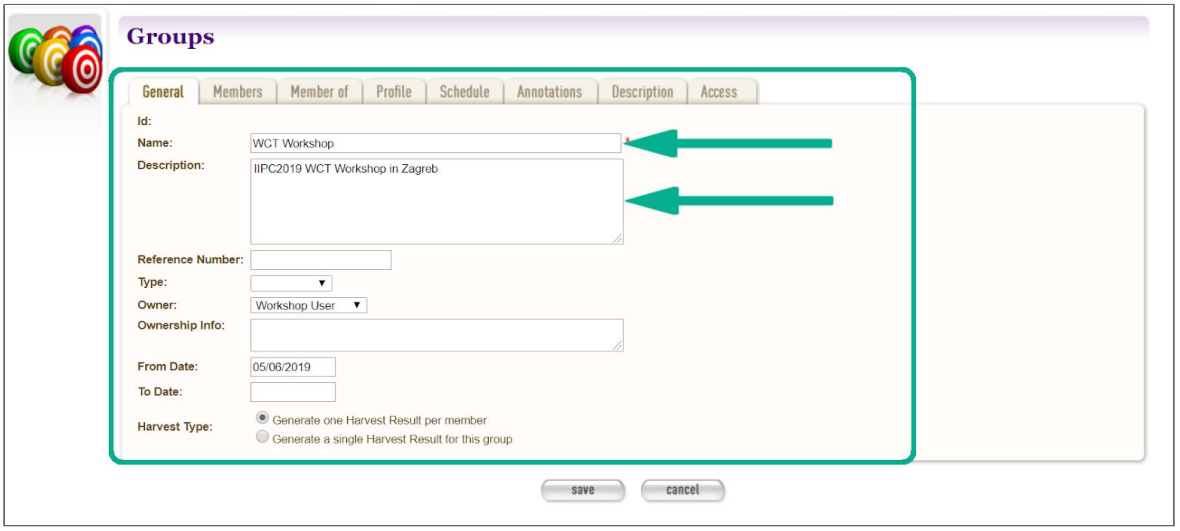
Groups

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.

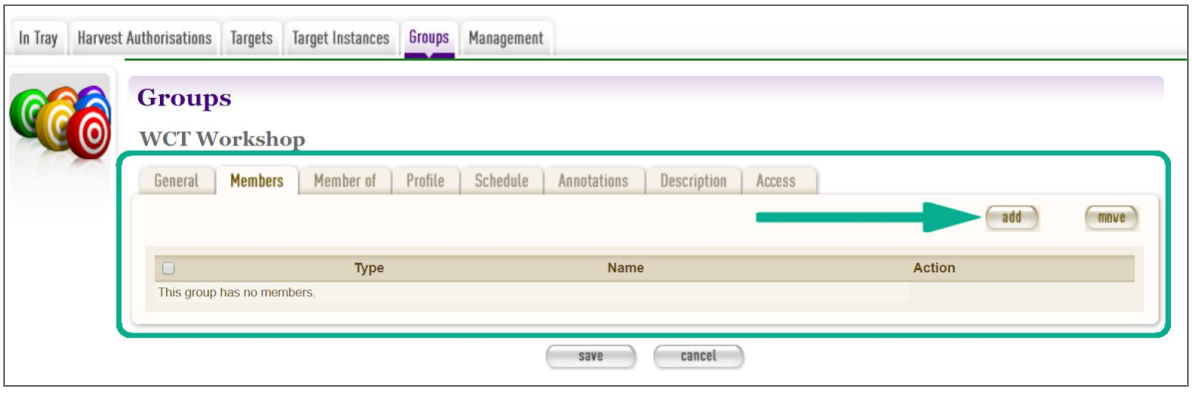
1. Navigate to the Groups tab within WCT.



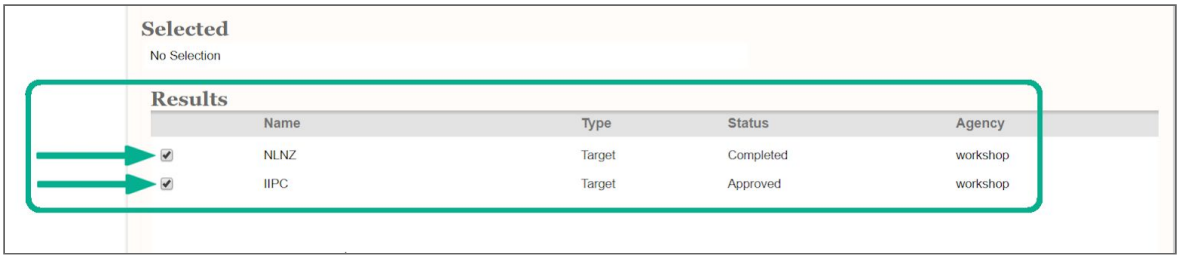
2. Click the Create New button
3. In the General tab, enter the following detail
   Name: WCT Workshop
   Description: IIPC2019 WCT Workshop in Zagreb



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



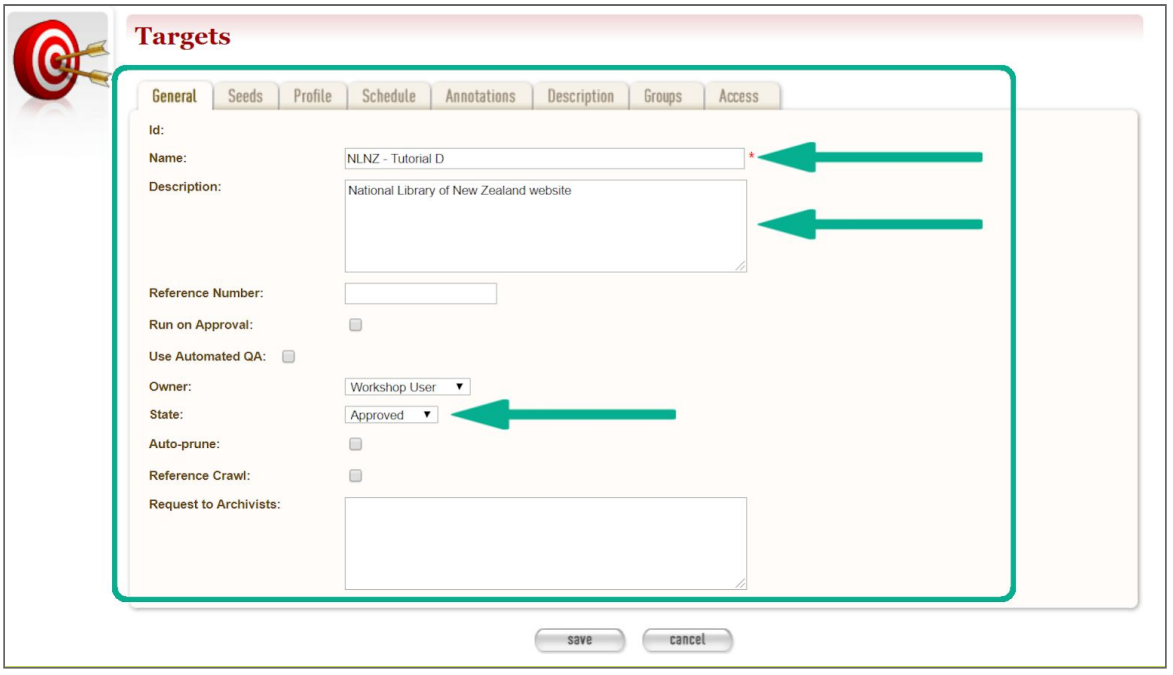
5. Tick the checkboxes next to all the Targets you have created during this workshop. Click Save to assign them to this group.



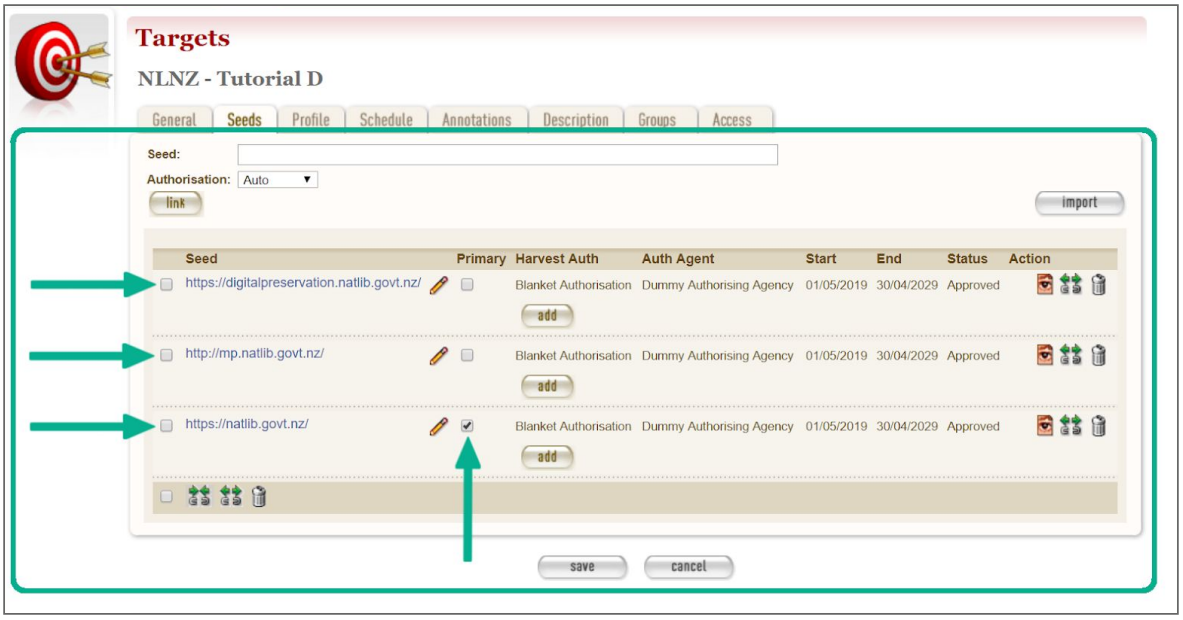
Setup Target

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.

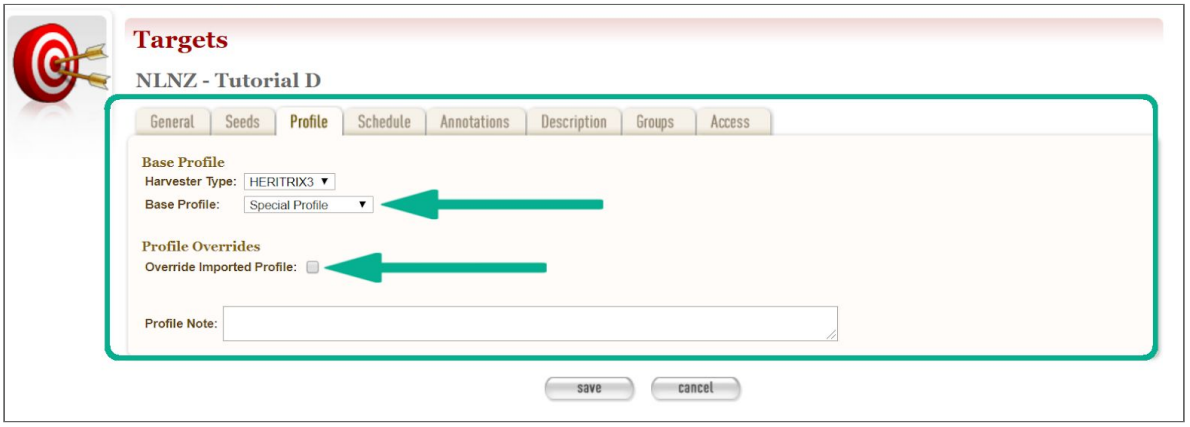
1. Navigate to the Targets tab in WCT.
2. Create a new Target.
3. In General tab, enter the details for a new website you want to harvest.
   Name: <short name of your institution> + " - Tutorial D"
   Description: <full name of your institution> website
   State: Approved



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.

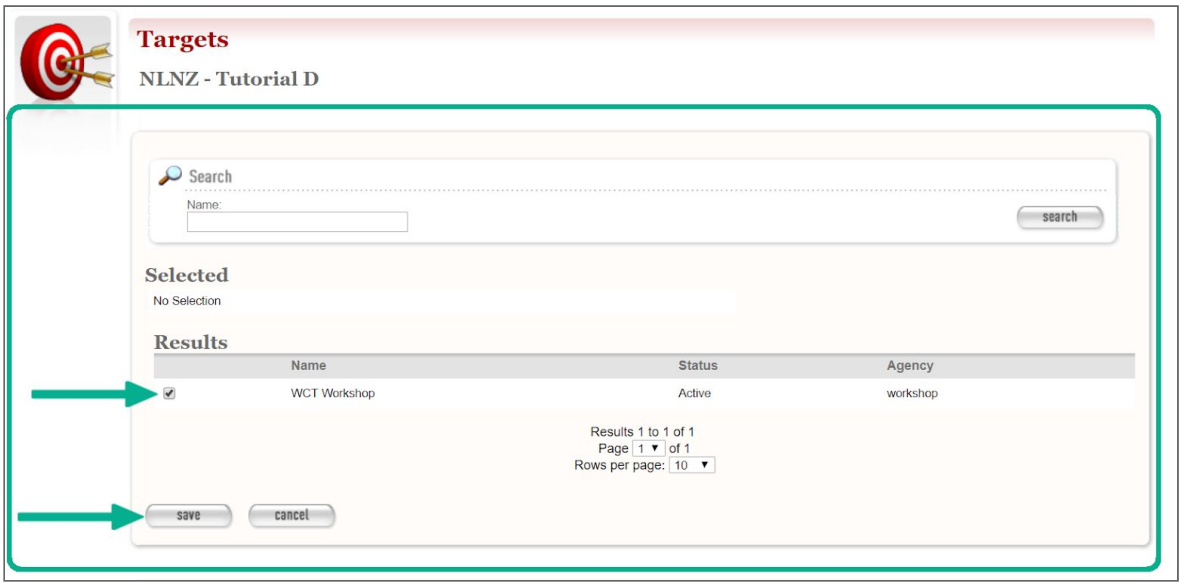


5. Navigate to the Profile tab, and select the new imported profile “Special Profile”. *Ticking the Override Imported Profile checkbox will display the xml editor for making Target specific changes to the profile.*



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

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

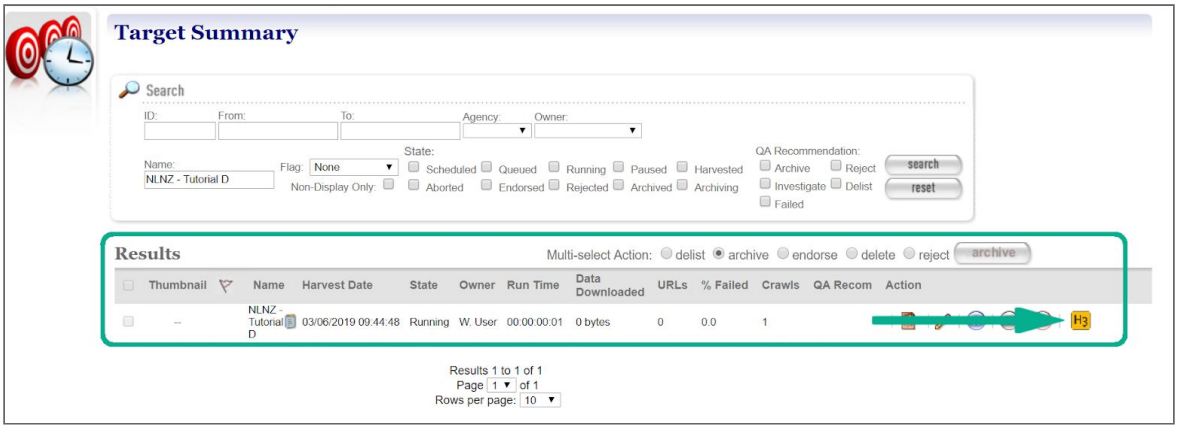


8. Navigate to the Schedule tab, tick the Harvest Now box, and click Save.
9. Click the View Target Instances icon next to the Target.
10. Start the Target Instance using the Harvest Now option.

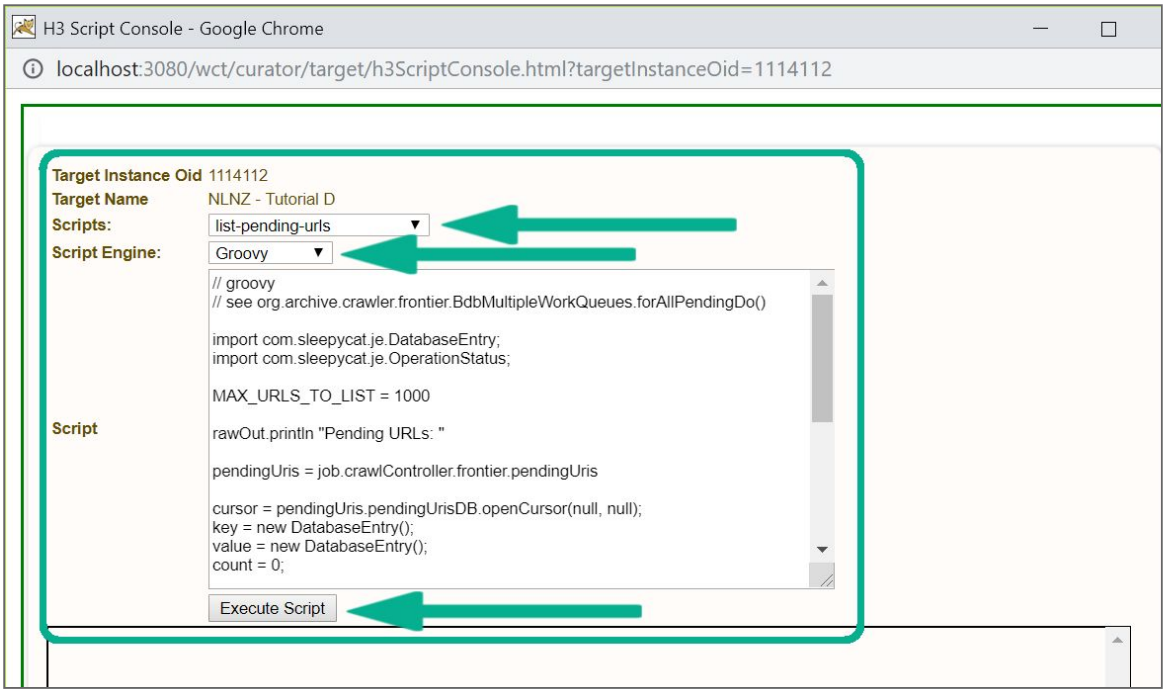
Heritrix 3 Scripting Console

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

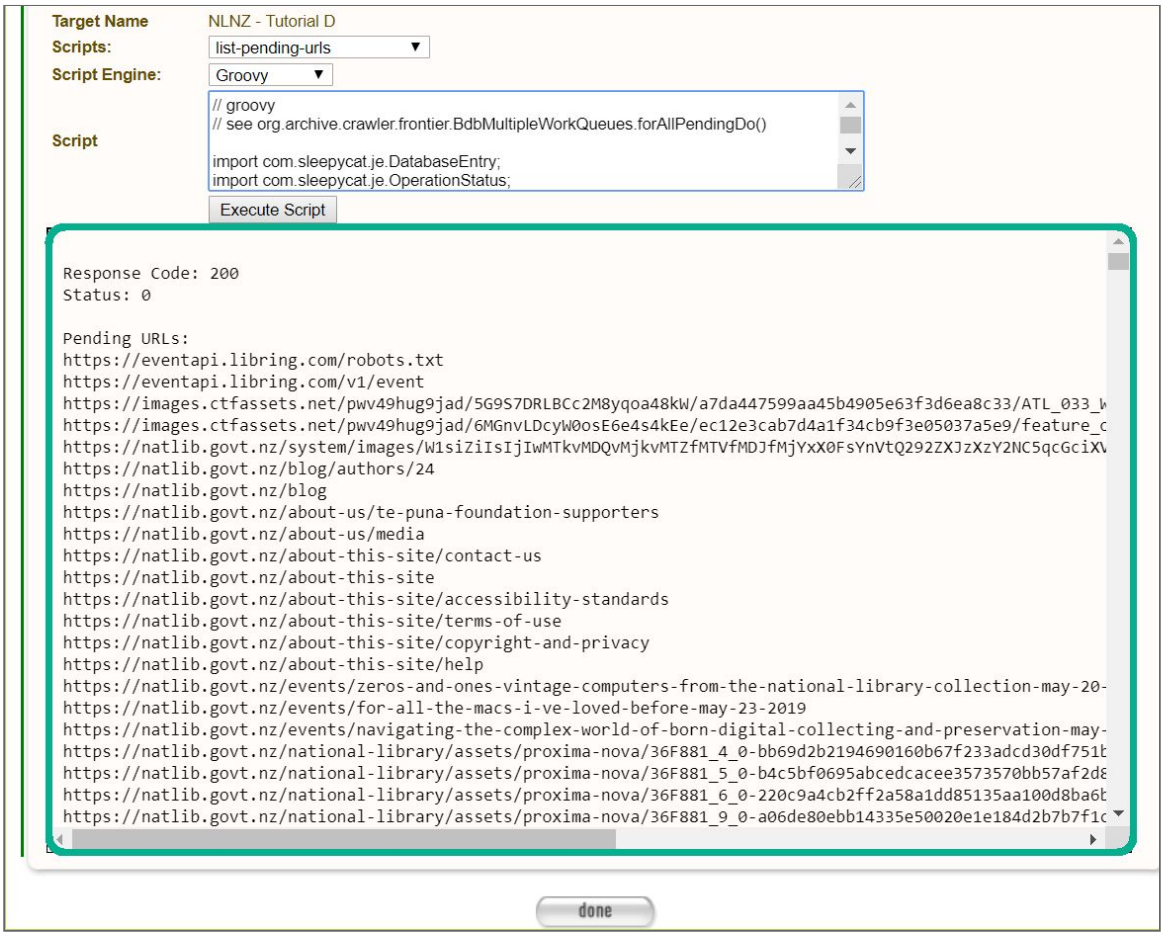
1. Observe the scheduled Target Instance move into a Running state. Click the H3 Script Console icon.



2. Here we can edit and execute predefined and custom scripts against the running crawl. Enter the following details and click the Execute Script button.
   Scripts: list-pending-urls
   Script Engine: Groovy



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

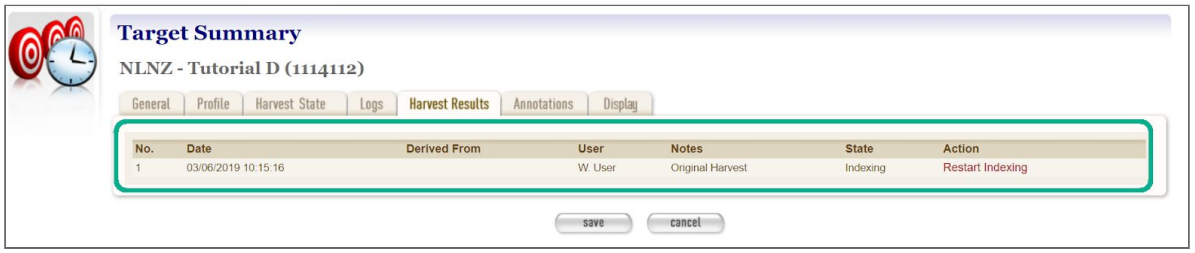


4. Feel free to experiment with the other two available scripts.
    - The *empty-frontier* script will remove all remaining URLs from the crawl queue.
    - The *remove-host-from-frontier* script will remove all URLs containing a specified hostname from the crawl queue. That hostname is set in the script.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

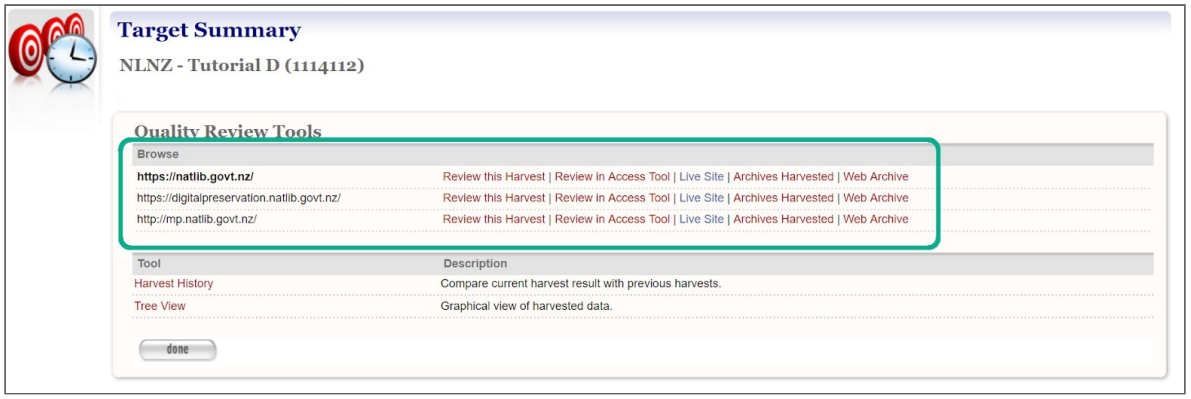
Quality review

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

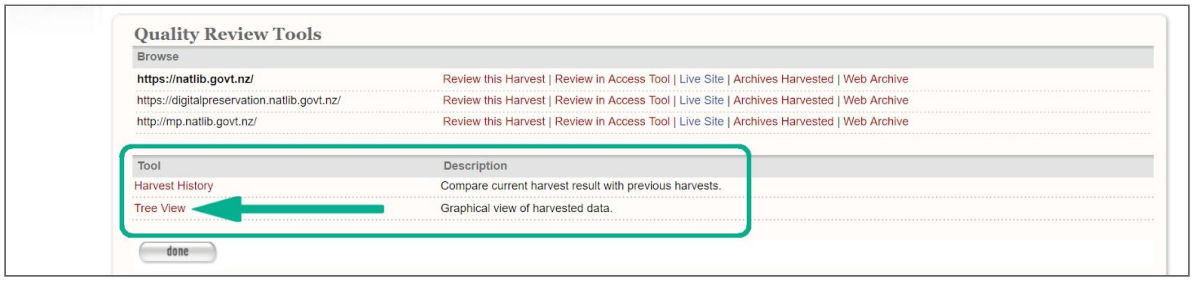
1. Navigate back to the Target Instance tab.
2. Stop the running Target Instance.
3. Once it has finished and moved into the Harvested state, click the Edit icon.
4. Navigate to the Harvest Results tab and wait for the indexing to complete.



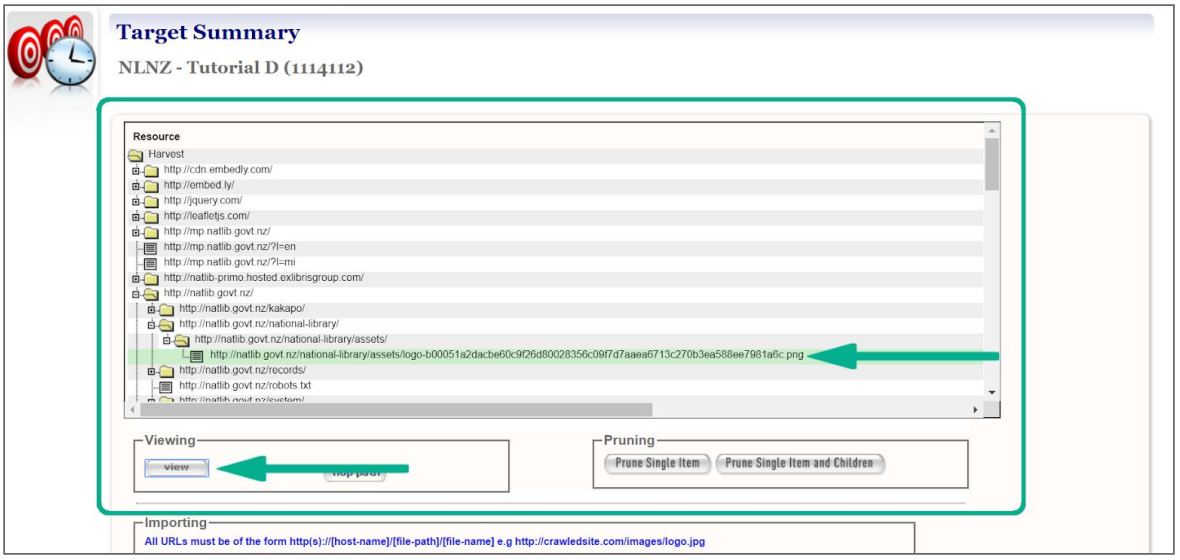
5. Once the Harvest Result has been indexed, click the Review link.
6. Notice the multiple seed URLs listed. These can each be viewed directly in OpenWayback by clicking the *Review in Access Tool* link.



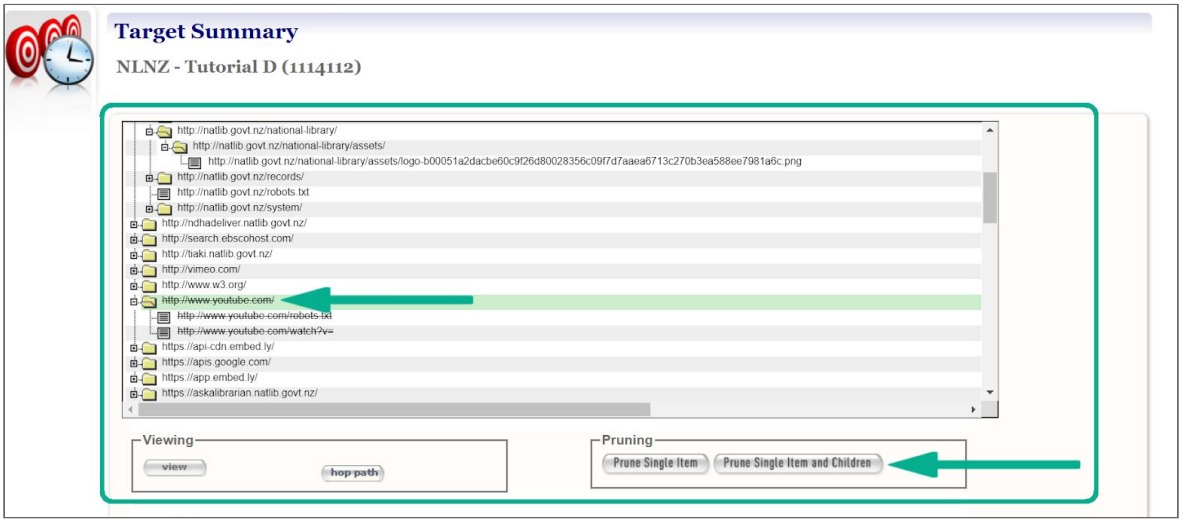
7. Click the *Tree View* link



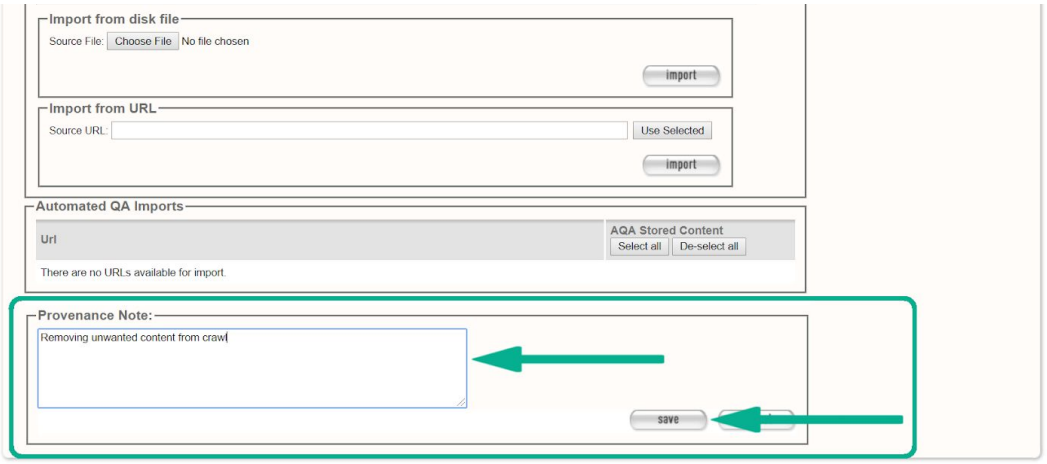
8. All the harvested resources can be viewed within a tree structure. Highlight an individual resource and click the *View* button to show the document directly using OpenWayback.



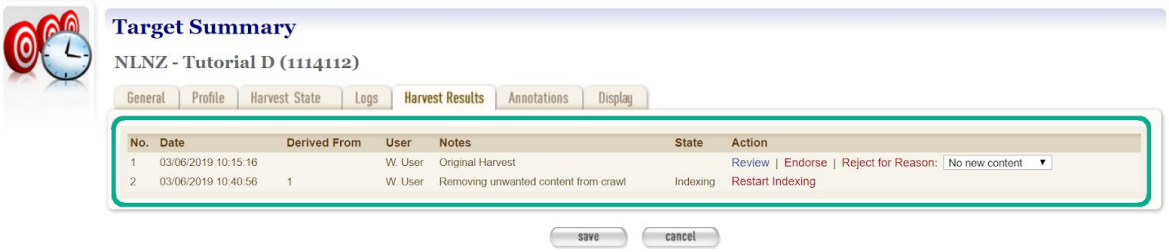
9. The tree view can be used to prune unwanted URLs from a particular hostname or sub-path. Highlight a folder in the tree view, and click the *Prune Single Item and Children* button.



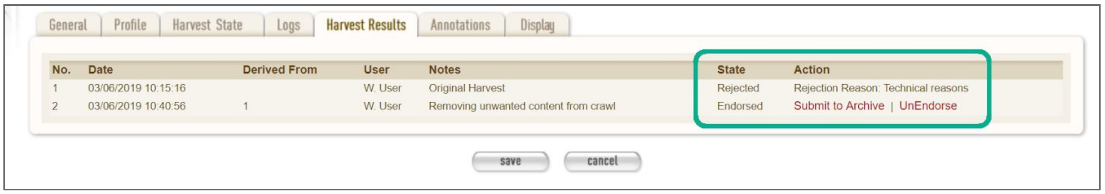
10. The resources with the ~~strike through~~ are now staged for pruning. Enter the following details and click Save.
    Provenance Note: Removing unwanted content from crawl.



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.



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



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.

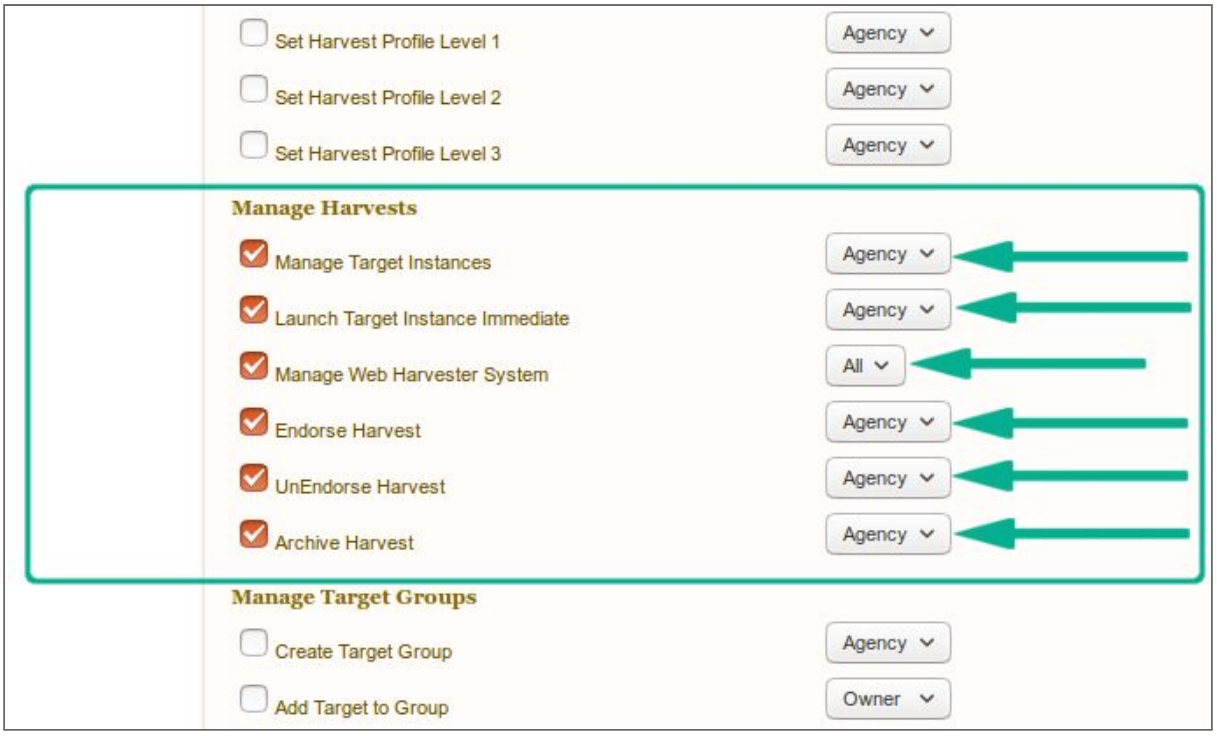
In this tutorial we will create a new user with restricted permissions who will be assigned specific tasks within the WCT workflow. Let's suppose someone in your organisation is tasked with quality control and endorsement of completed harvests. We will create a user with the corresponding privileges.

Creating the role

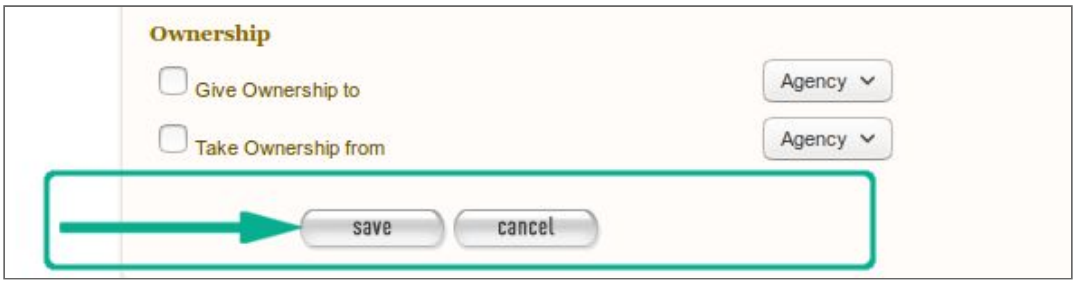
1. Click on the add new button behind Roles in the Management tab.

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

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



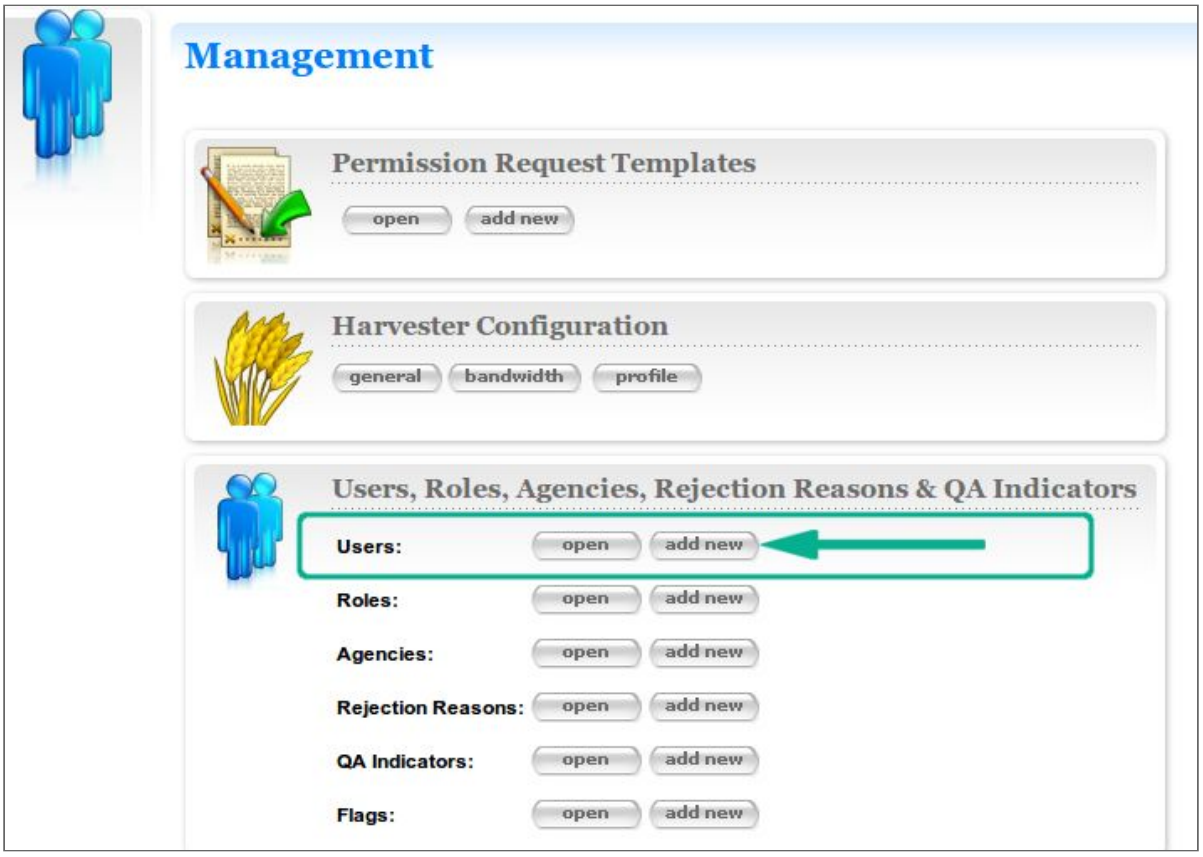
7. Save the role.



Create the user

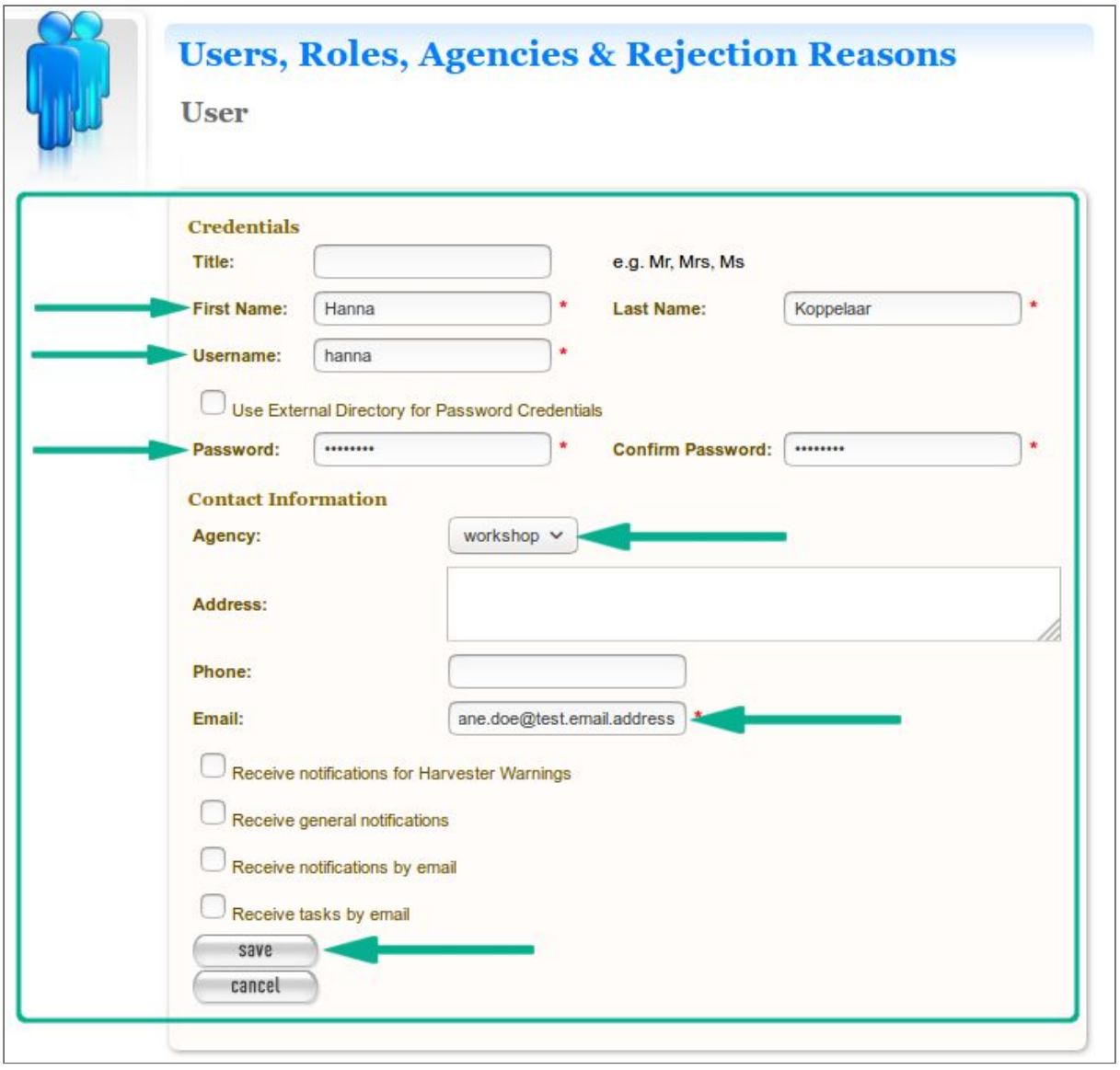
Now, let's create a user with this role.

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

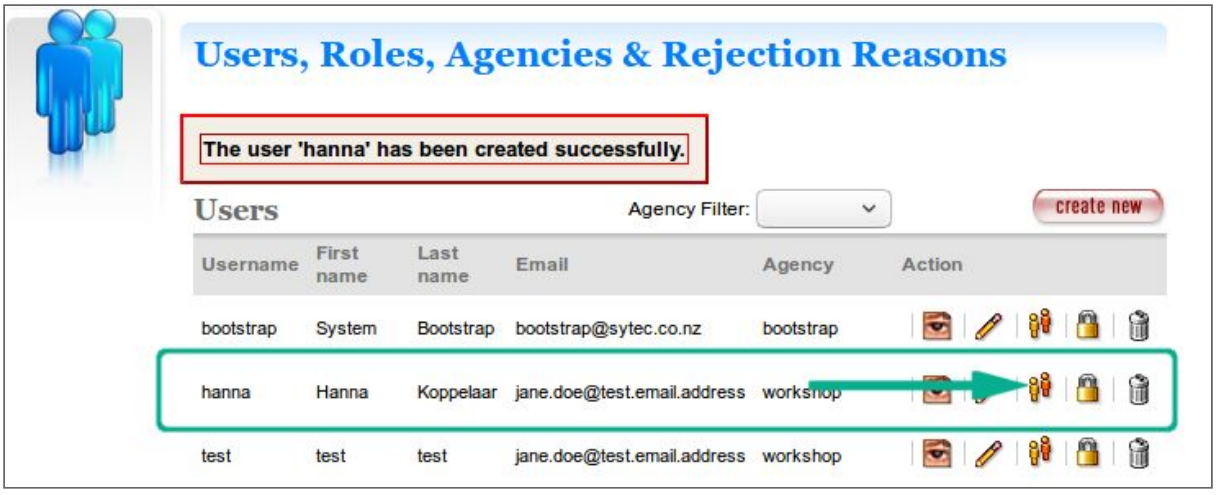


2. Use your own name in First Name, Last Name and Username.
3. Enter and confirm a password.
   This is a temporary password. The first time a user logs in, they are forced to set their own new password.
4. Select your agency.
5. In the Email field, just put: jane.doe@test.email.address.

6. Save the user.

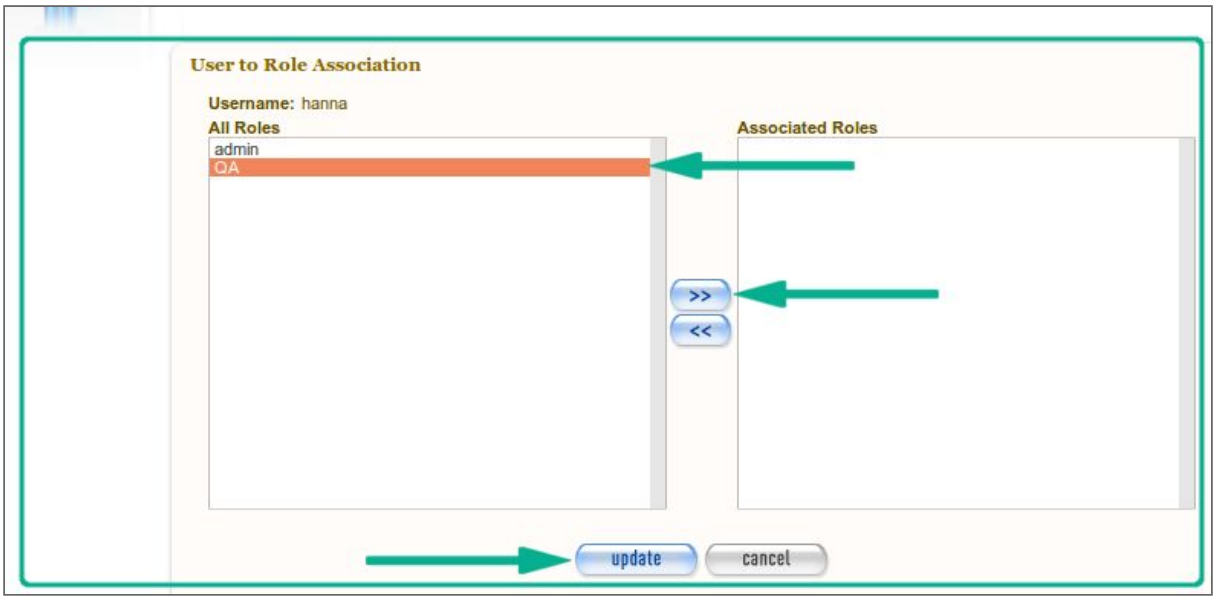


7. Open the group memberships screen for the new user.



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

9. Click on the update button.



10. Logout.

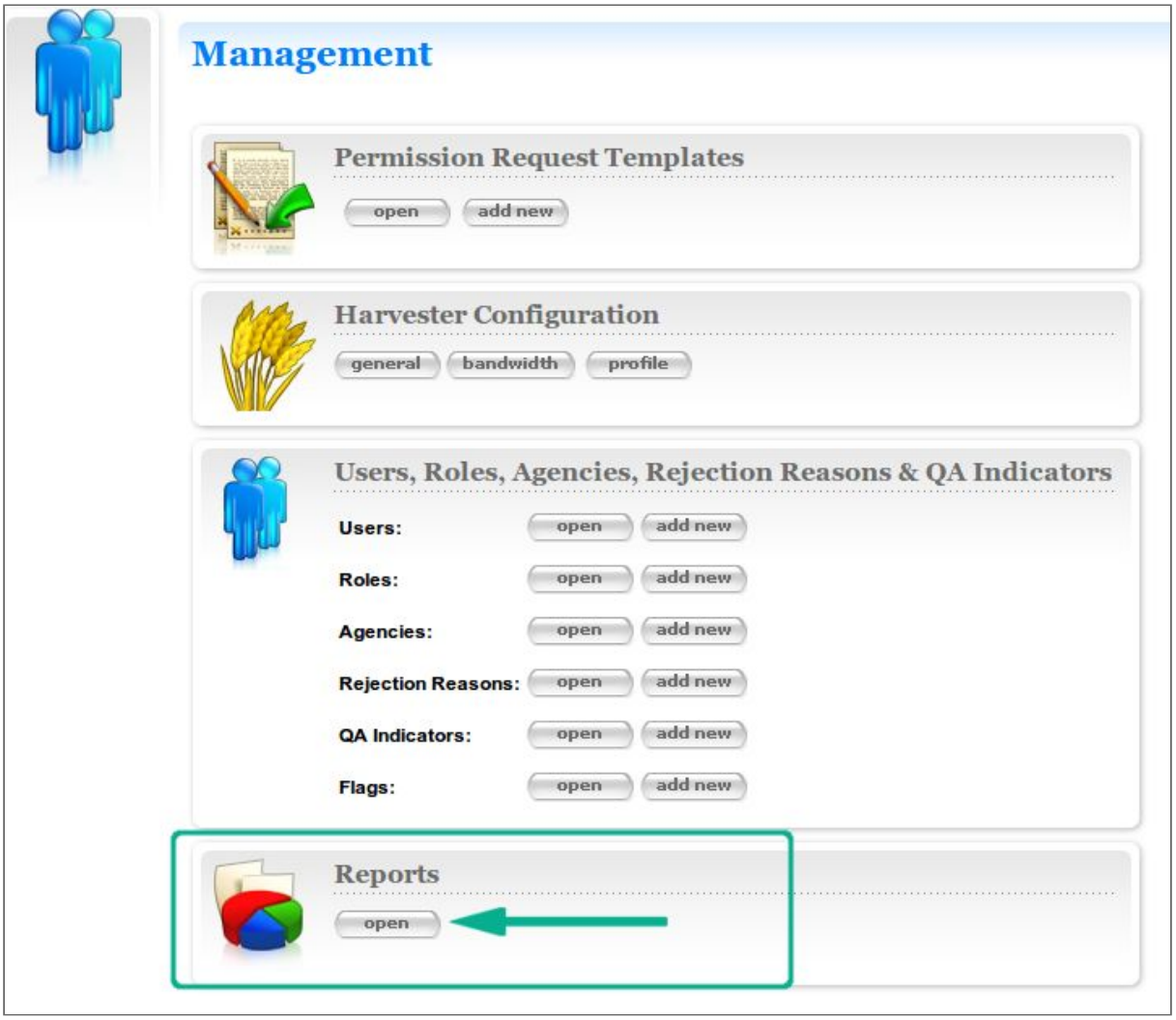
11. Now login as the new user you've just created. You will be asked to change your password, since this is your first login. Notice that the interface now offers less options, because this user has less privileges. E.g. when you open the Targets screen, you will not see the button that allows you to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

Tutorial F - Reporting

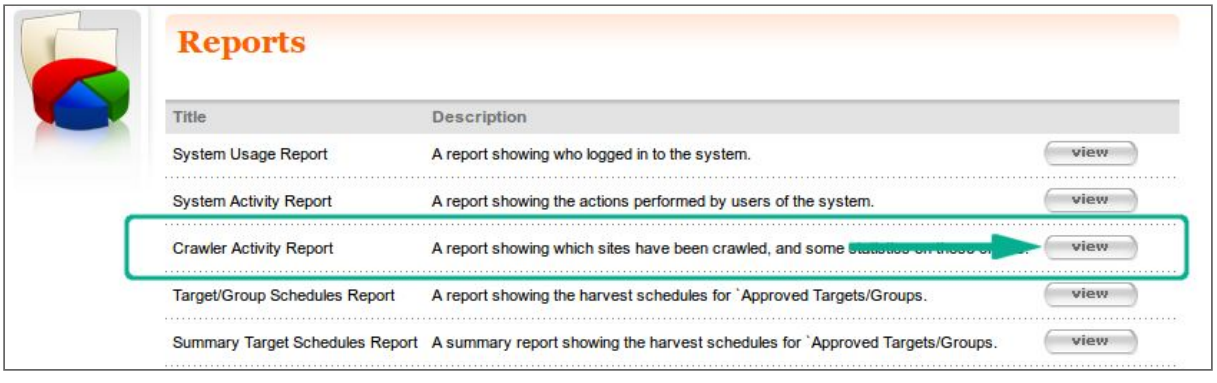
What you will learn [reporting]

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

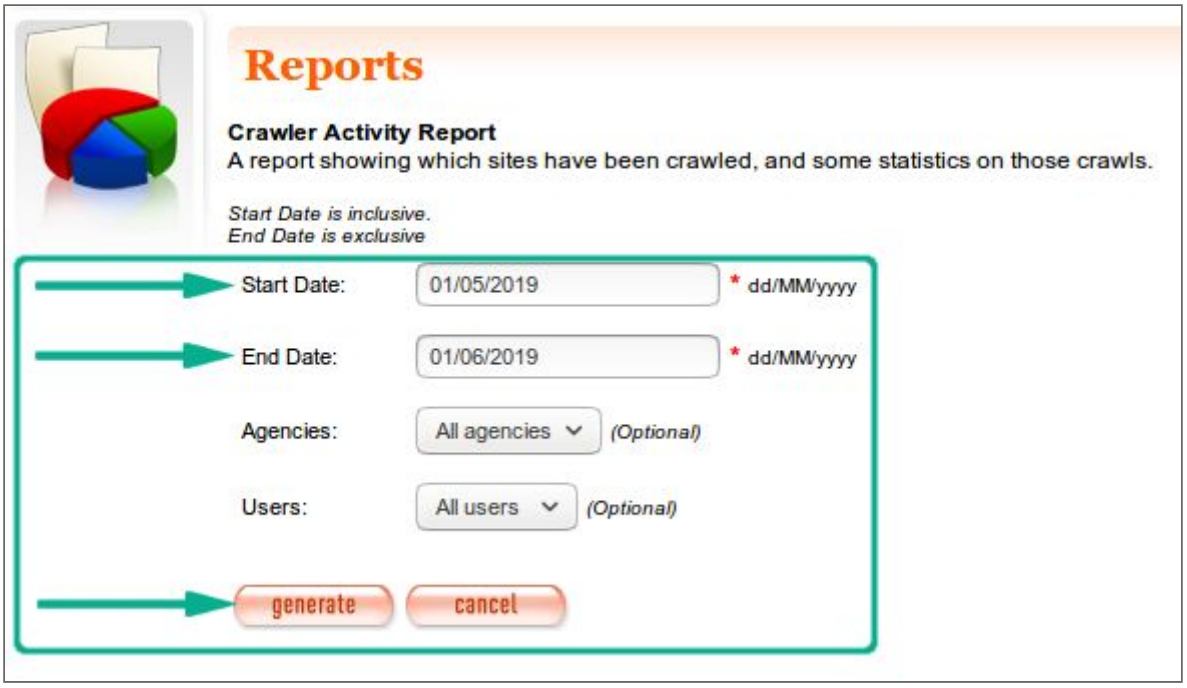
1. Click on the open button under Reports in the Management tab.



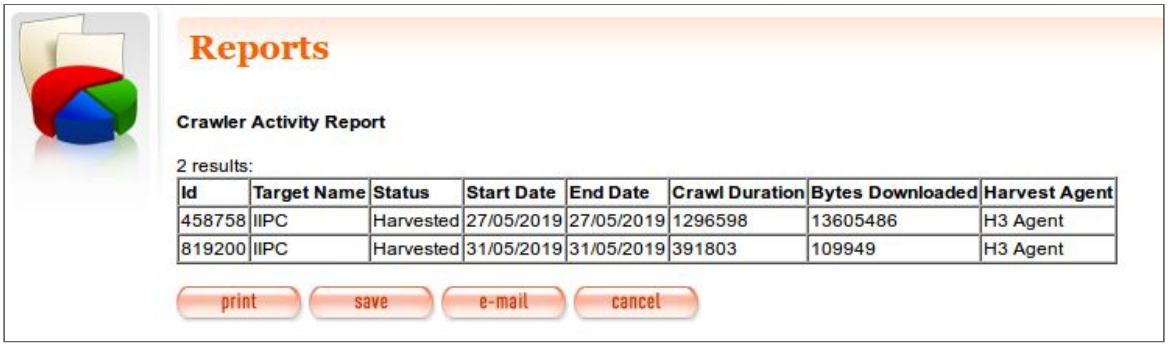
2. Click on the view button behind the Crawler Activity Report.



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



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.*



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