



WEB CURATOR TOOL

Quick Start Guide

Version 1.6.2

13 November 2015



Contents




Introduction	3
Basic functions: authorisation, targets, harvests	3
Terminology	3
General workflow	5
Contents of this document	7
Home Page	8
In Tray	10
Harvest Authorisations	11
Sample harvest authorisation.....	11
Permission status	11
Harvest Authorisation page	12
To create a harvest authorisation request:.....	13
Enter general information about the request	13
Enter URLs you want to harvest.....	14
Enter agencies who grant permission	14
Enter details of permissions requested	15
Generate a letter to send to the authorising agent	16
Targets.....	18
Target status.....	18
Targets page.....	19
To create a target:	20
Enter general information about the target.....	21
Enter the sites you want to harvest	21
Enter a schedule for the target	21
Target Instances.....	23
Target instance status.....	23
Target instance page	24
To review target instances:	26
Review harvest results.....	28
Endorse, reject, un-endorse or archive harvest results	30
Appendix A: Detailed Workflow	31



Introduction












Basic functions: authorisation, targets, harvests

The Web Curator Tool facilitates retrieving and archiving web pages across the internet. It includes the following basic functions:

-  setting up requests to harvest web pages —
see [To create a harvest authorisation request: page 13](#)
-  setting up target schedules for harvesting —
see [To create a target: page 20](#)
-  reviewing harvested content —
see [To review target instances: page 26](#).

Terminology

Some basic terms used with the Web Curator Tool include:

-  **harvest** — the process of crawling the web and retrieving specific web pages; can also refer to the files retrieved
-  **authorisation** — approval for you to harvest, save, and provide access to published web material
-  **permission** — within an authorisation, specific record of an authorisation, including authorising agencies, the dates during which permissions apply and any restrictions on harvesting or access
-  **authorising agency** — a person or organisation who authorises a harvest; often a web site owner or copyright holder
-  **target** — defines the portion of the web you want to harvest (such as a web site or a set of web pages), with crawler configuration details and a schedule of harvest dates
-  **target instance** — a single harvest of a target, scheduled to occur at a specific date and time
-  **seed** — a starting URL for a harvest, such as the root address of a web site. A harvest usually starts with a seed and includes all pages underneath that seed in the website
-  **indicator** — a quality assurance metric used to quantify the success of a harvest (eg: the amount of content downloaded)
-  **recommendation** — the advice obtained by using one or more indicators to determine the if a harvest successfully captured the content from a website
-  **automated QA** — the automated quality assurance process that runs after a harvest completes that provides a recommendation
-  **flag** — an arbitrary group created and assigned to one or more target instances



reference crawl — a target instance that has been archived and marked as a baseline to which all future harvests will be compared for a specific target

General workflow

The general workflow for the Web Creator Tool is to:

- 1** Create [Harvest Authorisations](#) (page 11), which include
 - **URL patterns** (what you want to harvest)
 - **authorising agencies** (who grant permission for the harvest)
 - **permissions** (requests for an authorising agency to approve specific harvests of one or more URL patterns).
- 2** For each permission, create a task to manage the approval process.
- 3** Claim the approval task from the [In Tray](#) (page 10), and:
 - create a **permission request letter**
 - email or print and post the letter to the authorising agency
 - change the permission's status from 'pending' to 'requested'.
- 4** When you receive a response from the authorising agency and:
 - edit the permission record, for example adding any special conditions
 - change its status to 'approved' or 'rejected'.
- 5** Create [Targets](#) (page 18) that defines the web material you want to harvest, technical harvest parameters and schedules for harvesting.
- 6** After harvests run, review [Target Instances](#) (page 23) and:
 - prune the results as needed
 - endorse or reject the results
 - archive endorsed results.

The following diagram illustrates the general flow of authorisations, targets, and harvests:

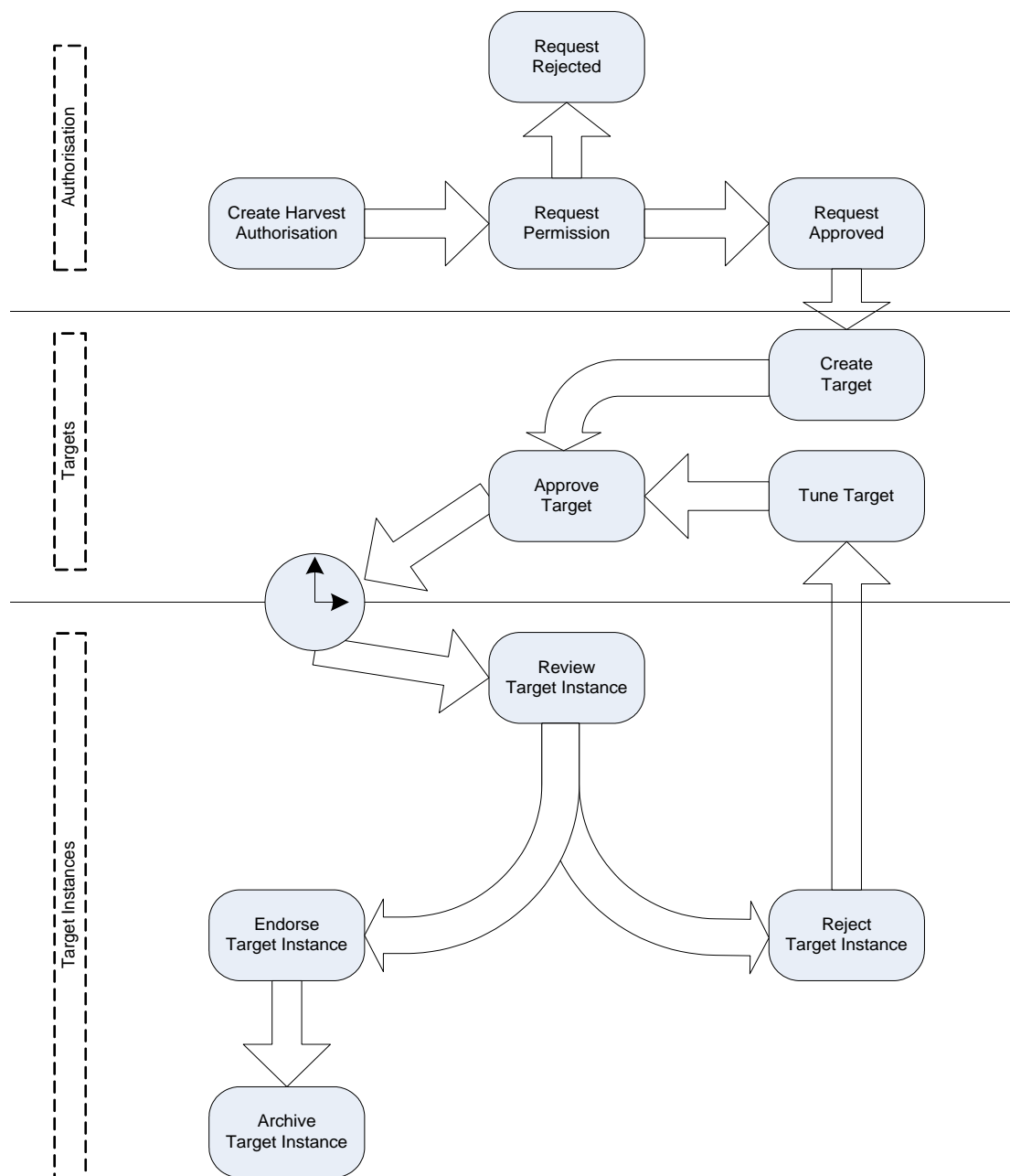








Figure 1. Web Curator Tool process flow

For a more detailed flowchart, see [Appendix A: Detailed Workflow](#), page 31.

Contents of this document

The Web Curator Tool Quick Start Guide includes the following sections:

-  **Home Page** (page 8) — an overview of the Web Curator Tool home page and summary of each of the major functions
-  **In Tray** (page 10)— an overview of the In Tray, which displays tasks and notifications for the logged-in user
-  **Harvest Authorisations** (page 11) — procedures for adding and editing requests for permission to harvest web pages
-  **Targets** (page 18) — procedures for adding and editing schedules for harvesting web pages
-  **Target Instances** (page 23) — procedures for adding, editing, reviewing, and archiving particular harvests
-  **Appendix A: Detailed Workflow** (page 31) — flowchart detailing the complete Web Curator Tool process.

Home Page

The **Web Curator Tool Home Page** includes the following functions:

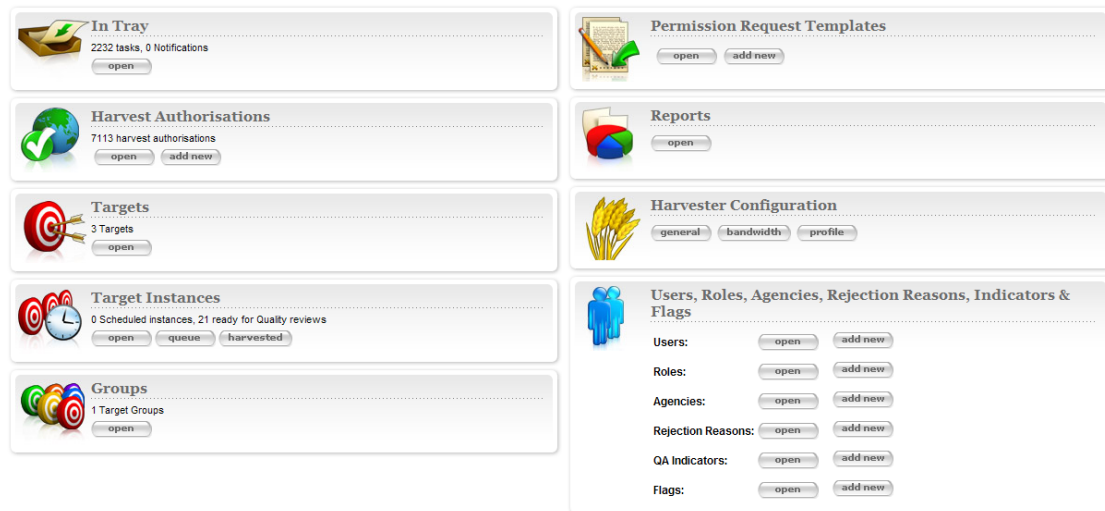











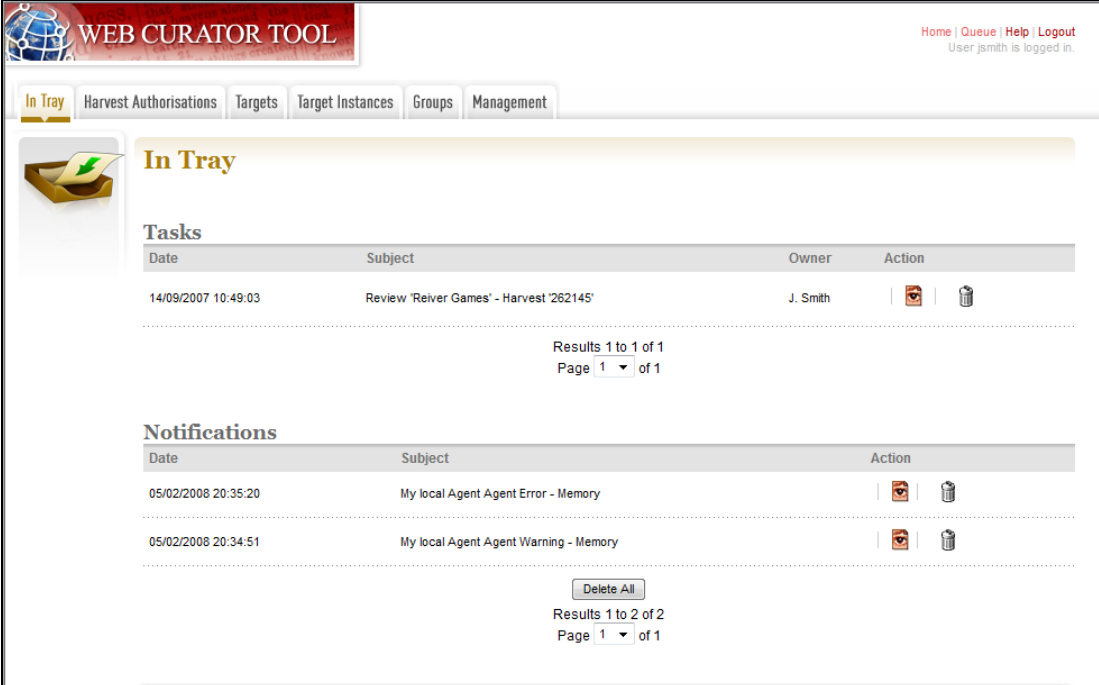
Figure 2. Home Page

-  **In Tray** — view tasks that require action and notifications that display information, specific to the user
-  **Harvest Authorisations** — create and manage harvest authorisation requests
-  **Harvest Configuration** — *system administrator function* to configure time-based bandwidth restrictions (how much content can be downloaded during different times of the day or week) and harvest profiles (such as how many documents to download, whether to compress them, delays to accommodate the hosting server, etc.)
-  **Reports** — *system administrator function* to generate reports on system activity
-  **Permission Request Templates** — create templates for permission request letters
-  **Targets** — create and manage targets and their schedules
-  **Target Instances** — view the harvests scheduled in the future and review the harvests that are complete
-  **Groups** — create and manage collections of targets, for collating meta-information or harvesting together
-  **Users, Roles & Agencies, Rejection Reasons, Indicators & Flags** — *system administrator function* to create and manage users, agencies, roles, privileges, rejections reasons, QA indicators and flags

*The functions that display on the **Web Curator Tool Home Page** depend on the user's privileges.*

In Tray

The **In Tray** displays *Tasks* and *Notifications* specific to your login.





WEB CURATOR TOOL

Home | Queue | Help | Logout
User jsmith is logged in.





In Tray

Tasks

Date	Subject	Owner	Action
14/09/2007 10:49:03	Review 'Reiver Games' - Harvest '262145'	J. Smith	 

Results 1 to 1 of 1
Page 1 of 1

Notifications

Date	Subject	Action
05/02/2008 20:35:20	My local Agent Agent Error - Memory	 
05/02/2008 20:34:51	My local Agent Agent Warning - Memory	 

Delete All





Results 1 to 2 of 2
Page 1 of 1

Figure 3. In Tray

Tasks are events that require action from you (or others with your privileges), for example endorsing or archiving a harvest.

Notifications are information such as system messages.

For each listing, you can:

-  — **View** details of the task or notification
-  — **Delete** the task or notification
-  — **Claim** the task (for example, if you are among those who can endorse a harvest, you can claim the task so that you can then perform the endorsement).
-  — **Un-claim** the task (for example, if you have accidentally claimed a task that is more appropriately carried out by someone else then you can release the task back to the pool of un-claimed tasks for someone else to claim).

Note that the **In Tray** — and each **Web Curator Tool** page — has tabs across the top to access the main system functions, which match the icons on the [Home Page](#).



Harvest Authorisations

Before you can harvest, archive, or display a set of web pages, you must get permission from the owner(s). The Web Curator Tool helps you do this using **harvest authorisation records**. Each harvest authorisation record is a collection of related URL patterns, authorising agencies, and permissions.

Sample harvest authorisation

For example, to harvest web pages from 'The Alphabet Soup Company', you might create a harvest authorisation record called 'Alphabet Soup'. This would include:



URL patterns to cover the company's three web sites:

- <http://www.alphabetsoup.com/>*
- <http://www2.alphabetsoup.com/>*
- <http://extranet.alphabetsoup.com/>*



authorising agencies for the two organisations responsible for updating content on these sites:

- The Alphabet Soup Company
- Food Incorporated.



permissions, linking each authorising agency with one or more URL patterns, and optionally specifying a time period and any special conditions or access restrictions (such as 'only users from New Zealand can view archived content'); for example:

- The Alphabet Soup Company to approve restriction-free access, on an open-ended basis, to <http://www.alphabetsoup.com/>* and <http://www2.alphabetsoup.com/>*
- Food Incorporated to approve NZ-only access, for the period 1/1/2006 through 31/12/2006, to <http://www.alphabetsoup.com/>* and <http://www2.alphabetsoup.com/>*

Permission status

Each permission request has a status:



pending — the permission has been created, but not yet assigned to a user for sending the request letter



requested — a request for permission has been sent to the authorising agency



approved — the authorising agency has approved the permission



rejected — the authorising agency has refused the permission.

Harvest Authorisation page

The **Harvest Authorisation** page lets you create and manage requests for permission to harvest web pages.

WEB CURATOR TOOL

Home | Queue | Help | Logout
User jsmith is logged in.

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

Harvest Authorisations

Search

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

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

search reset

Results [create new](#)

Id	Name	Auth Agent	Order No	Status	Action
267982	(Not) Alan Milburn	(Not) Alan Milburn		Pending, Approved	
295064	@teb: Answers in Technology for Expanding Business	@teb: Answers in Technology for Expanding Business		Approved	
265784	24 Hour Museum	24 Hour Museum		Approved	
280013	247 Media Network Limited	247 Media Network Limited		Approved	
287538	24dash	24dash		Approved(2)	
291818	25% ME Group	25% ME Group		Approved	

Figure 4. Harvest Authorisations

At the top of the page are:

- fields to search for existing harvest authorisation records by **ID**, **Name**, **Authorising Agent**, **Order Number**, **Agency**, **URL Pattern**, **Permissions File Reference** and/or **Permissions Status**
- a button to **create new** harvest authorisation requests.

Below that are search results. For each harvest authorisation record found, you can:

- **View** details
- **Edit** details
- **Copy** (and modify), for example if you are creating multiple, similar requests
- **Generate** a permission request letter.

Note that, as of release 1.3 of the software, all search pages that present the search results in a 'page at a time' fashion have been modified so that the user can elect to change the default page size from 10 to 20, or 50 or even 100! The user's preference will be remembered across sessions in a cookie.

To create a harvest authorisation request:





From the [Harvest Authorisations](#) page,

- 1 Click **create new**.

The **Create/Edit Harvest Authorisations** page displays:

Figure 5. Create/Edit Harvest Authorisations

The **Create/Edit Harvest Authorisations** page includes four tabs for adding or editing information on a harvest authorisation record:

-  **General** — general information about the request, such as a name, description and any notes
-  **URL Patterns** — patterns of URLs for which you are seeking authorisation
-  **Authorising Agencies** — the persons and/or organisations from whom you are requesting authorisation
-  **Permissions** — details of the authorisation, such as dates and status.

Enter general information about the request

- 2 On the **General** tab, enter basic information about the authorisation request.

The system will validate your entries and let you know if you leave out any required information.

- 3 To add a note (annotation) to the record, enter it and click **add**.

Enter URLs you want to harvest

- 4 Click the **URL Patterns** tab.

The **URL Patterns** tab includes a box for adding URL patterns and a list of added patterns.

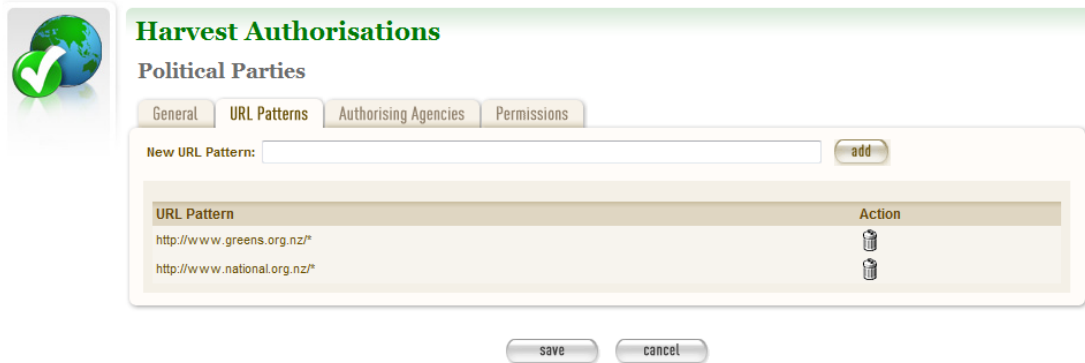


Figure 6. URL Patterns tab

- 5 Enter a pattern for the URLs you are seeking permission to harvest, and click **add**. Repeat for additional patterns.

You can use the wildcard * at the start of the domain or end of the resource to match the permission to multiple URLs. For example:

- `http://*.govt.nz/*` — to include all NZ Government sites
- `http://*.nz/*` — to include all sites in the *.nz domain space (supports a permission based on government legislation)
- `http://www.alphabetsoup.com/*` — to include all resources within the Alphabet Soup site (a standard permission granted directly by a company)
- `http://www.alphabetsoup.com/resource/*` — to include only the resources within the 'resource' section of the Alphabet Soup site (for the company to be more restrictive; granting for example just URL patterns resources/*, about/*, help/* and excluding other sections they do not want harvested).
- `http://*.alphabetsoup.com/*` — to include all resources on all sub sites of the specified domain.

Enter agencies who grant permission

- 6 Click the **Authorising Agencies** tab.

The **Authorising Agencies** tab includes a list of authorising agencies and buttons to search for or create new agencies.



Figure 7. Authorising Agencies tab

- 7** To add a new agency, click **create new**.
The **Create/Edit Agency** page displays.

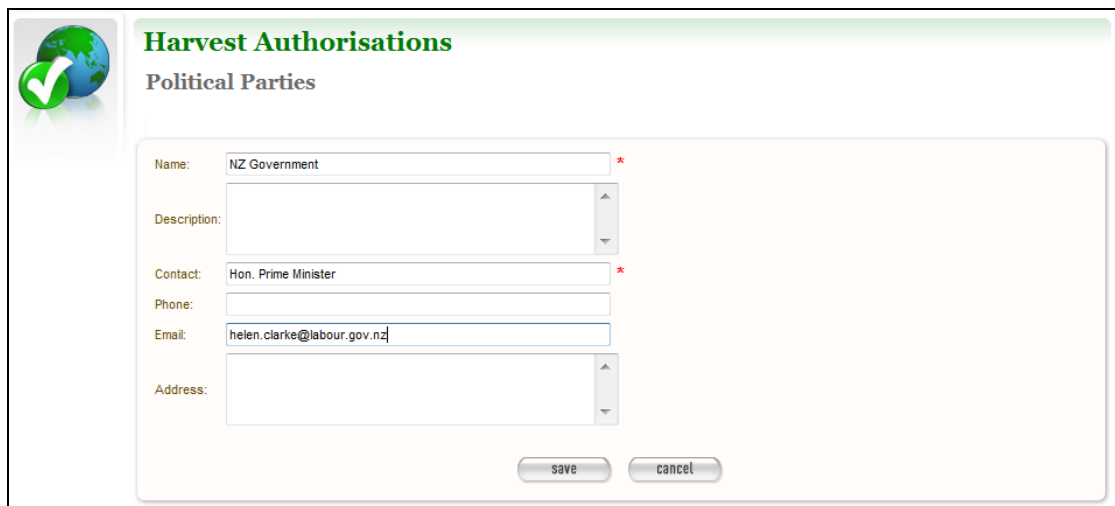


Figure 8. Create/Edit Agency

- 8** Enter the name, description, and contact information for the agency; and click **Save**.
The [Authorising Agencies tab](#) redisplay, showing the added agency.

Enter details of permissions requested

- 9** Click the **Permissions** tab.
The **Permissions** tab includes a list of permissions requested showing the status, agent, dates, and URL pattern for each.



Status	Authorising Agent	From	To	URL Patterns	Action
Pending	NZ Government	01/01/2008	01/01/2009	http://www.national.org.nz/* http://www.greens.org.nz/*	   

Figure 9. Permissions tab

- 10** To add a new permission, click **create new**.
The **Create/Edit Permission** page displays.

Harvest Authorisations
Political Parties

Authorising Agent: NZ Government

Dates: 01/01/2007 * to 01/01/2008 dd/mm/yyyy

Status: Pending

Special Restrictions:

Copyright Statement: These sites are not bound by crown copyright

Copyright URL:

Access Status: Open (unrestricted) access

Open Access Date:

Quick Pick:

Display Name:

Uris: ☐ http://www.greens.org.nz/* * ☒ http://www.national.org.nz/* *

File Reference:

Assign Approval Task: No

Exclusions

URL	Reason
No exclusions have been defined.	

Annotations

Date	User	Notes	Action
There are no annotations available.			

save cancel

Figure 10. Create/Edit Permission

- 11** Select an agent, enter the dates you want to harvest, tick the URL patterns you want to harvest, enter special restrictions, etc.; and click **Save**.


The [Permissions tab](#) redisplay, showing the added permission.

- 12** Click **Save** to save the harvest authorisation request.

The new (or changed) record displays on the General tab of the [Create/ Edit Harvest Authorisations](#) page.

After adding or editing a harvest authorisation record, you must save before clicking another main function tab (eg, Targets or Groups), or your entries will be lost.

Generate a letter to send to the authorising agent

- 13** Click  next to the harvest authorisation request.

The system generates and displays the letter.

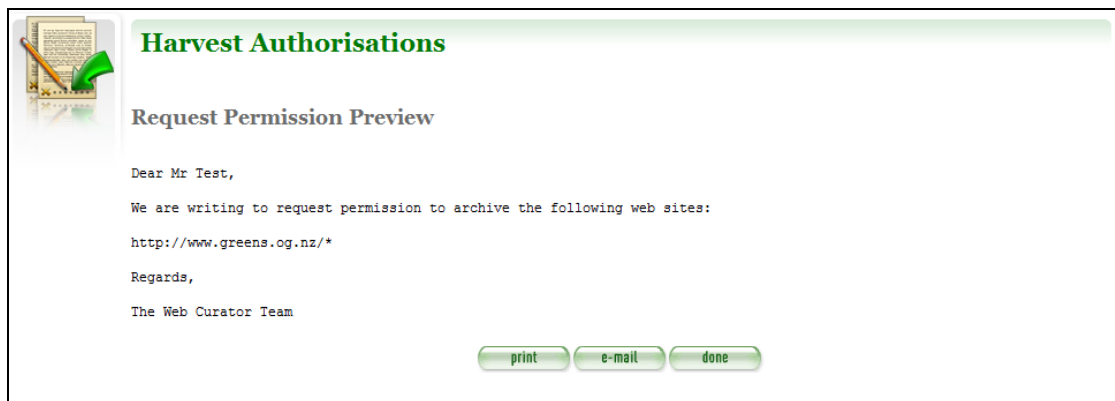


Figure 11. Permission Request Letter

- 14** Click to **print** or **e-mail** the letter to the agent.

The system sends the letter and changes the permission status to 'requested'.

- 15** Click **Done**.

The [Harvest Authorisations](#) page redisplays.










Targets

Once you have received authorisation to harvest a set of web pages, you must create a **target**, which defines exactly when and what the Web Curator Tool will retrieve.

Targets include URL patterns, profiles, and schedules for harvesting.

Target status

Each target also has a status:

-  **pending** — a work in progress, not ready for approval
-  **nominated** — ready for approval, displaying as a task in the [In Tray](#) of all users with privileges to approve this target
-  **rejected** — rejected by the approver; may be issues with specific permissions or a decision not to harvest this target
-  **approved** — ready for harvest
-  **complete** — harvested; all schedules associated with the target completed
-  **cancelled** — harvest scheduling was cancelled before completed
-  **reinstated** — the target was reinstated from the complete, cancelled, or rejected state; but is not yet ready for approval (at which time it will be put into the nominated state).

Targets page

You manage targets using the **Targets** page:

Targets

Search

ID: Name: Seed: Agency: The British Library User: John Smith

Member of: State: ☐ Pending ☐ Reinstated ☐ Nominated ☐ Rejected ☐ Approved ☐ Cancelled ☐ Completed

search reset

Results [create new](#)

Id	Name	Agency	Owner	Status	Seeds	Action
293507	68 Dean Street	The British Library	John Smith	Approved	http://www.sixty8.com	
291657	ARC	The British Library	John Smith	Approved	http://arc.co.uk/	

Results 1 to 2 of 2
Page 1 of 1

In Tray | Harvest Authorisations | Targets | Groups | Target Instances | Reports | Management

Figure 12. Targets

At the top of the page are:

- fields to search for existing targets by **ID**, **Name**, **Seed** (root URL of a Web site), **Agency**, **User**, **Member of** and **State**
- a button to **create new** targets.

Below that are search results (defaults to show targets that you own). For each target found, you can:

- **View** details
- **Edit** details
- **Copy** (and modify), for example if you are creating multiple, similar targets.

To create a target:

From the [Targets](#) page,

- 1 Click **create new**.

The **Create/Edit Targets** page displays.

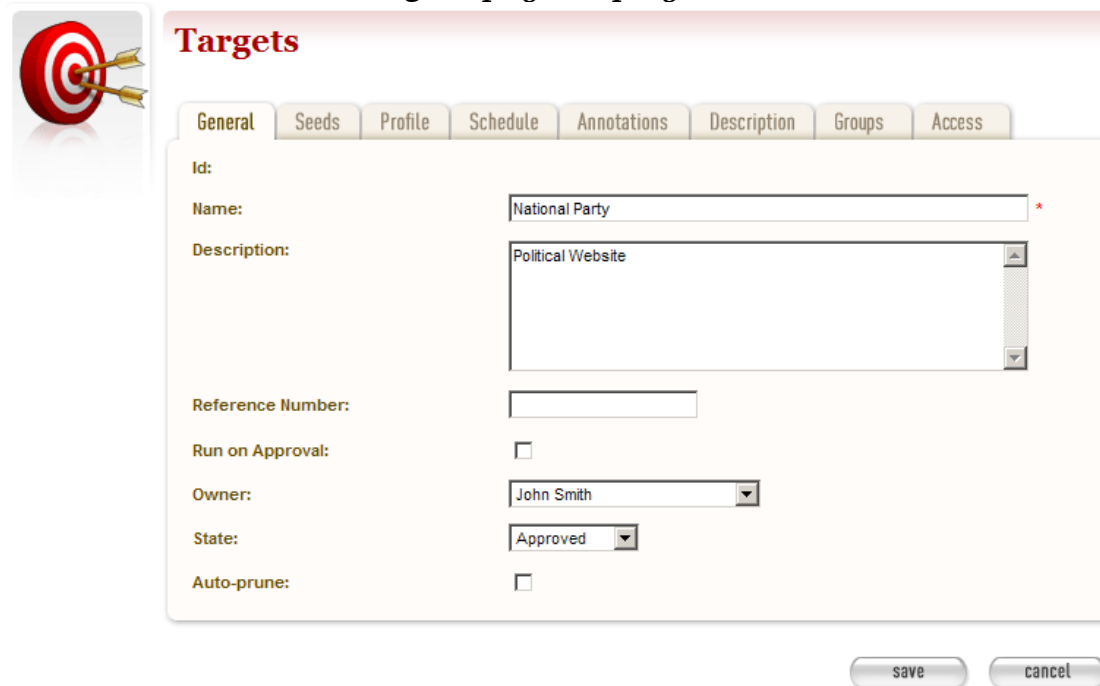










Figure 13. Create/Edit Targets

The **Create/Edit Targets** page includes five tabs for adding or editing information about targets:

-  **General** — general information about the target, such as a name, description, owner, and status
Enabling the **Auto-prune** checkbox causes WCT to identify pruned items from the last archived harvest and prunes those items from subsequent harvests
-  **Seeds** — base URLs for web sites to harvest
-  **Profile** — technical instructions on how to harvest (entered by system administrator)
-  **Schedule** — dates and times to perform the harvest
-  **Annotations** — notes about the target.
-  **Description** — meta data about the target
-  **Groups** — groups that the target is a member of
-  **Access** — meta data defining access to the target once archived

Enter general information about the target

- 2 On the **General** tab, enter basic information about the target.

The system will validate your entries and let you know if you leave out any required information.

Enter the sites you want to harvest

- 3 Click the **Seeds** tab.

*The **Seeds** tab includes a box for adding the base URL of each web site you want to harvest and list of previously added seeds.*

The screenshot shows the 'Targets' application window. The title bar says 'Targets'. Below it, the target name 'National Party' is displayed. There are several tabs: 'General', 'Seeds', 'Profile', 'Schedule', 'Annotations', 'Description', 'Groups', and 'Access'. The 'Seeds' tab is active. It contains a 'Seed:' text box, an 'Authorisation:' dropdown menu set to 'Auto', a 'link' button, and an 'import' button. Below these is a table with columns: 'Seed', 'Primary', 'Harvest Auth', 'Auth Agent', 'Start', 'End', 'Status', and 'Action'. The table has one row with the seed 'http://www.national.org.nz/', which is marked as 'Primary' (checked), 'Harvest Auth' (Political Parties), 'Auth Agent' (NZ Government), 'Start' (01/01/2008), 'End' (01/01/2009), and 'Status' (Approved (expired)). There are icons for adding, unlinking, and deleting seeds. At the bottom of the window are 'save' and 'cancel' buttons.

Figure 14. Seeds tab

- 4 Enter the root URL of a web site for this target.
- 5 Select an **Authorisation** for the target:
- **Auto** finds all harvest authorisation records whose URLs match the seed
 - **Add Later** enters the seed unlinked to any permissions, which can be added later
 - **Quick Pick** enters seeds that do not need individual permissions, for example where government legislation covers a large number of seeds.
- 6 Click **link**. Repeat for additional sites.

The seed displays in the list (see above).

*You can also use the **Import** button to import a precompiled list of seeds.*

The multiple selection bar at the bottom of the list allows you to link, unlink and delete multiple selected seeds.

Enter a schedule for the target

- 7 Click the **Schedule** tab.

The **Schedule** tab includes a list of schedules and a button to create a new schedule.

Figure 15. Schedule tab

8 Click **create new**.

The **Create/Edit Schedule** page displays fields for entering a schedule.

Figure 16. Create/Edit Schedule

- 9** Enter **From** and **To** dates for when the harvest will run; select a **Type** of schedule, eg 'Every Monday at 9:00pm' or 'Custom' — if you select 'Custom', enter details of the schedule; and click **Save**.
- 10** Click **save** at the bottom of the page to save the target.

After adding or editing a target record, you must save before clicking another main function tab (eg, Harvest Authorisation or Groups), or your entries will be lost.

*You can also add general notes about the target by clicking the **Annotations** tab.*













Target Instances

Target Instances are actual dates and times a specific Target is run. For example, a target might specify that particular websites should be harvested every Monday at 9pm; a target instance would be the actual harvest run at 9pm on Monday 24 July 2006.

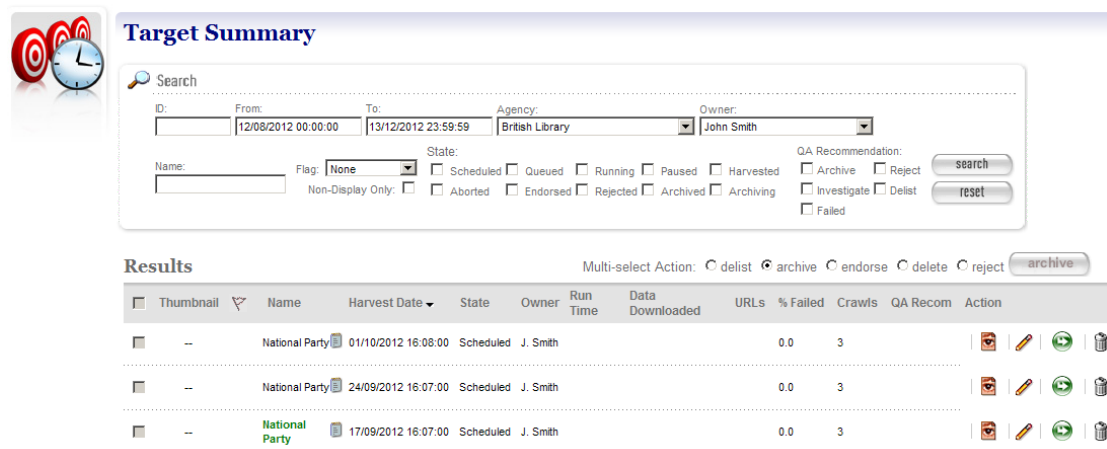
Target instance status

Each target instance has a status:

-  **scheduled** — waiting for its scheduled time
-  **queued** — reached its scheduled time, but cannot run immediately; eg, not enough bandwidth available or the available harvest agents have reached their maximum concurrent harvest count
-  **running** — in the process of harvesting
-  **stopping** — finished harvesting, performing final clean-up
-  **paused** — paused during harvesting
-  **aborted** — manually aborted; deleted any collected data
-  **harvested** — completed or stopped; data collected is available for review
-  **endorsed** — harvested data reviewed and deemed suitable for archiving
-  **rejected** — harvested data reviewed and found not suitable for archiving (ie, content is incomplete or not required)
-  **archived** — harvested content submitted to the archive.

Target instance page

You manage target instances from the **Target Instance** page:



The screenshot shows the 'Target Summary' page. At the top left is a logo with two target icons. The main section is a search form with fields for ID, From (12/08/2012 00:00:00), To (13/12/2012 23:59:59), Agency (British Library), and Owner (John Smith). There are also checkboxes for Name, Flag (None), State (Scheduled, Queued, Running, Paused, Harvested, Aborted, Endorsed, Rejected, Archived, Archiving, Failed), and QA Recommendation (Archive, Reject, Investigate, Delist). Search and reset buttons are present. Below the search form is a 'Results' section with a table of target instances. The table has columns: Thumbnail, Name, Harvest Date, State, Owner, Run Time, Data Downloaded, URLs, % Failed, Crawls, QA Recom, and Action. There are three rows of data, all for 'National Party' and 'J. Smith'. The first two rows are 'Scheduled' and the third is 'Harvested'. Each row has a set of action icons in the 'Action' column.






Thumbnail	Name	Harvest Date	State	Owner	Run Time	Data Downloaded	URLs	% Failed	Crawls	QA Recom	Action
--	National Party	01/10/2012 16:08:00	Scheduled	J. Smith				0.0	3		[Icons]
--	National Party	24/09/2012 16:07:00	Scheduled	J. Smith				0.0	3		[Icons]
--	National Party	17/09/2012 16:07:00	Scheduled	J. Smith				0.0	3		[Icons]

Figure 17. Target Instances

At the top of the page are:

 fields to search for existing target instances by **ID**, by **From** and **To** dates, **Agency**, **Owner**, **Name**, **State** and **QA Recom**.

Below are search results. For each target instance found, you can:

-  — **View** details
-  — **Edit** details
-  — **Harvest Now** start harvesting this Target Instance immediately
-  — **Delete** a scheduled target instance so its harvest does not run (*target instances can only be deleted in the 'scheduled' or 'queued' states. Target instances in the 'queued' state may only be deleted if harvesting has not previously started*).
-  — **Target Annotation:** displays any annotations defined for this target instance's target.

Operations on multiple target instances can be performed using the **Multi-select Action** radio button. Note that the target instance checkbox will be enabled only for those target instances in a valid state for the selected multi-select action:

delist: cancels all future schedules for the selected target instances.











endorse: endorses the selected target instances.

archive: archives the selected target instances.

delete: deletes all selected target instances in a valid state (eg: scheduled target instances).

reject: when selected, a rejection reason drop-down box is displayed and clicking the action button will reject the selected target instances with the selected rejection reason:

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

<input type="checkbox"/> Thumbnail	Name	Harvest Date	State	Owner	Run Time	Data Downloaded	URLs	% Failed	Crawls	QA Recom	Action
	National Party	13/09/2012 15:14:28	Archived	J. Smith	00:00:00:18	521.11 KB	34	0.0	3	Reject	 
	National Party	13/09/2012 16:16:46	Harvested	J. Smith	00:00:00:18	521.1 KB	34	0.0	3	Reject	 
	National Party	24/09/2012 16:07:00	Scheduled	J. Smith				0.0	3		  

Sortable fields:

Harvest Date ▼

Clicking on the **Name**, **Harvest Date**, **State**, **Run Time**, **URLs**, **% Failed** or **Crawls** columns will sort the search results by that column.












Harvest Date ▲

Clicking the same column again will perform a reverse sort of the column

QA Recom

Hovering over the QA Recommendation will display a list of the three most recent harvest status and any annotations for the target instance:

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

<input type="checkbox"/> Thumbnail	Name	Harvest Date	State	Owner	Run Time	Data Downloaded	URLs	% Failed	Crawls	QA Recom	Action
	National Party	13/09/2012 15:14:28	Harvested	J. Smith	00:00:00:18	521.11 KB	34	0.0	3	Reject	 
	National Party	24/09/2012 16:07:00									  
	National Party	01/10/2012 16:08:00									  

Date	URLs	Data	Job Status	Status
13/09/2012 15:14:28	34	521.11 KB	Finished	Harvested
24/09/2012 16:07:00	34	521.11 KB	Finished	Harvested
01/10/2012 16:08:00	34	521.11 KB	Finished	Harvested

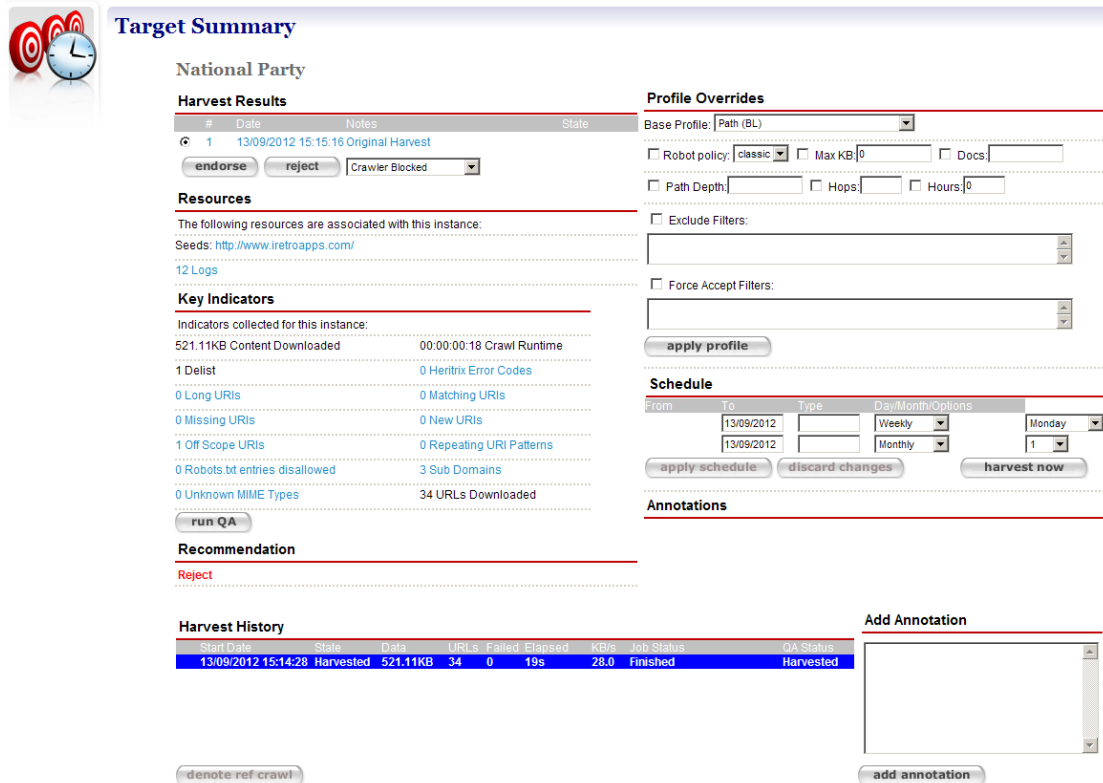
Annotations for the target instance:

- The crawler encountered a crawler trap. In most cases this can be resolved by applying filters and re-harvesting the site with refined crawl settings.
- The crawler has prevented the crawler from reaching significant parts of the site such as images or style sheets. Crawl settings can be refined to ignore robots.txt and the site re-gathered.
- Significant content is inaccessible. For example if the content is dependent on dynamic user interactions such as user input in search forms, Flash and Flash Navigation JavaScript menus.
- Streaming media content forms a significant part of the website.
- The content is dependent on embedded external applications e.g. Google Earth, Google Maps, YouTube etc.
- The crawler has gone beyond its scope, for example external links have been followed.
- Obscene or adult content appears on the website.

To review target instances:







- 1 Click the name of the target instance to view the target instance summary page.

The summary page is composed of panels that provide access to the QA Indicators and Recommendation, and draws together existing functionality into a single location.



The screenshot shows the 'Target Summary' page for the 'National Party' instance. The page is divided into several sections:

- Harvest Results:** A table showing harvest details. One entry is visible: #1, Date: 13/09/2012 15:15:16, Notes: Original Harvest, State: Crawler Blocked. Buttons for 'endorse', 'reject', and 'Crawler Blocked' are present.
- Resources:** A section for associated resources, including seeds (http://www.iretoapps.com/), logs (12 Logs), and a 'run QA' button.
- Key Indicators:** A section displaying various indicators collected for the instance, such as '521.11KB Content Downloaded', '00:00:00:18 Crawl Runtime', '1 Delist', '0 Long URIs', '0 Missing URIs', '1 Off Scope URIs', '0 Robots.txt entries disallowed', '0 Unknown MIME Types', '0 Heritrix Error Codes', '0 Matching URIs', '0 New URIs', '0 Repeating URI Patterns', '3 Sub Domains', and '34 URLs Downloaded'. A 'run QA' button is also present.
- Recommendation:** A section with a 'Reject' button.
- Profile Overrides:** A section for overriding the base profile (Path (BL)). It includes checkboxes for 'Robot policy' (classic), 'Max KB' (0), 'Docs' (0), 'Path Depth' (0), 'Hops' (0), 'Hours' (0), 'Exclude Filters', and 'Force Accept Filters'. An 'apply profile' button is at the bottom.
- Schedule:** A section for scheduling harvests. It includes fields for 'From' (13/09/2012), 'To' (13/09/2012), 'Type' (Weekly/Monthly), 'Day/Month/Options' (Monday/1), and buttons for 'apply schedule', 'discard changes', and 'harvest now'.
- Annotations:** A section for adding annotations, with an 'Add Annotation' button and a text area.
- Harvest History:** A table showing the history of harvests. One entry is visible: 13/09/2012 15:14:28, Harvested, 521.11KB, 34, 0, 19s, 28.0, Finished, Harvested.

-  **Harvest Results** — display the harvest results for the target instance; clicking the results displays the Harvest Results tab for the target instance
-  **Profile Overrides** — access to the base profile for the target instance
-  **Resources** — displays the seeds for the target instance; clicking a seed displays the Seeds tab for the target
-  **Schedule** — enables modification of existing schedules
-  **Key Indicators** — results of applying the Indicators defined in the System Administration Page for QA Indicators to the target instance; clicking a hyperlinked Indicator will display a generic report to explain the figure displayed. In the event that a target instance has been manually pruned, the **runQA** button is provided to re-compute the Indicator values and recommendation for the target instance.
-  **Annotations** — lists the notes about the target instance.



Recommendation — displays the final advice assigned to the target instance by considering all Indicator values. Hovering the mouse over the recommendation will display the advice for each indicator

Key Indicators	
Indicators collected for this instance:	
521.11KB Content Downloaded	00:00:00:18 Crawl Runtime
1 Dellist	0 Heritrix Error Codes
0 Long URIs	0 Matching URIs
0 Missing URIs	0 New URIs
1 Off Scope URIs	0 Repeating URI Patterns
0 Robots.txt entries disallowed	3 Sub Domains
0 Unknown MIME Types	34 URLs Downloaded
run QA	
Recommendation	
Reject	
Indicator	Advice Justification
Content Downloaded	None
Crawl Runtime	Reject The Crawl Runtime indicator value of 00:00:00:18 has fallen below its lower limit of 00:00:01:00
Dellist	None
Heritrix Error Codes	None
Long URIs	None
Matching URIs	None
Missing URIs	None
New URIs	None
Off Scope URIs	None
Repeating URI Patterns	None
Robots.txt entries	None



Add Annotation — enables notes for the target instance to be added.



Harvest History — displays all harvest history for the target instance's target. The current harvest is highlighted in blue. The harvest history for an archived target instance will be displayed with a radio option and clicking **denote ref crawl** will mark the selected archived target instance as the reference crawl for future crawls



Harvest History									
Start Date	State	Data	URLs	Failed	Elapsed	KB/s	Job Status	QA Status	
13/09/2012 16:16:46	Harvested	521.10KB	34	0	19s	28.0	Finished	Harvested	
<input checked="" type="radio"/> 13/09/2012 15:14:28	Archived	521.11KB	34	0	19s	28.0	Finished	Harvested	

[denote ref crawl](#)

When an archived target instance is denoted as a reference crawl, it is used as a baseline to compare the indicators for future crawls and is highlighted in red

Harvest History									
Start Date	State	Data	URLs	Failed	Elapsed	KB/s	Job Status	QA Status	
13/09/2012 16:16:46	Harvested	521.10KB	34	0	19s	28.0	Finished	Harvested	
<input checked="" type="radio"/> 13/09/2012 15:14:28	Archived	521.11KB	34	0	19s	28.0	Finished	Harvested	

2

Click  to view a target instance, or  to edit a target instance.

The **View/Edit Target Instance** page displays.








The image shows a web interface titled "Target Summary" for a target instance named "National Party (328007775)". On the left is a graphic of three red target icons and a clock. The main area has seven tabs: "General", "Profile", "Harvest State", "Logs", "Harvest Results", "Annotations", and "Display". The "General" tab is active, showing the following fields:

- Id:** 328007775
- Target Name:** National Party
- Schedule:** 17/09/2012 16:07:00
- Priority:** Normal (dropdown)
- Owner:** John Smith (dropdown)
- Agency:** British Library
- State:** Scheduled
- Bandwidth Percentage:** (empty input field)
- Flagged:** Reject Queue (dropdown)

At the bottom right are "save" and "cancel" buttons.

Figure 18. View/Edit Target Instance

The **View/Edit Target Instance** page includes seven tabs for viewing, running, or editing information about a target instance:

-  **General** — general information about the target instance, such as the target it belongs to, schedule, owner, agency, etc.
-  **Profile** — technical instructions on how to harvest (entered by system administrator), overriding those for the target as a whole
-  **State** — details of the harvest, for example total bandwidth and amount downloaded
-  **Logs** — files recording technical details of the harvest
-  **Harvest Results** — lists harvested content with options to review, endorse, reject, un-endorse and archive
-  **Annotations** — notes about the target instance.
-  **Display** — defines whether the harvested results should be displayed to the public.

- 3** To edit the instance, enter your changes and click **Save**.

Review harvest results

- 4** To manage harvest results (for a target instance that has run), click the **Harvest Results** tab.

A list of target results displays.

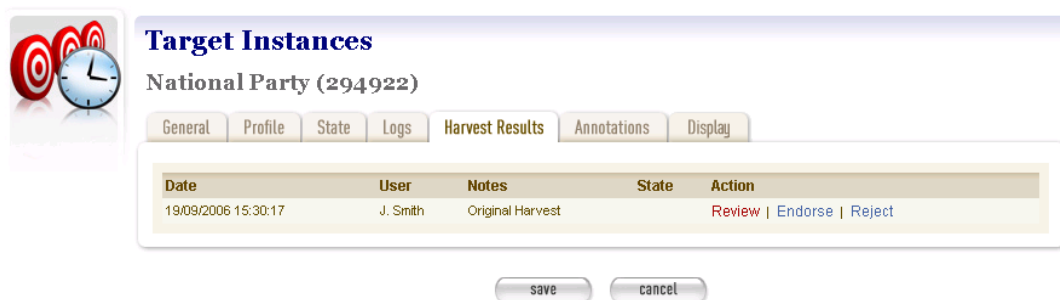


Figure 19. Harvest Results tab

- 5** To review a result, click **Review**.
Options for reviewing display.

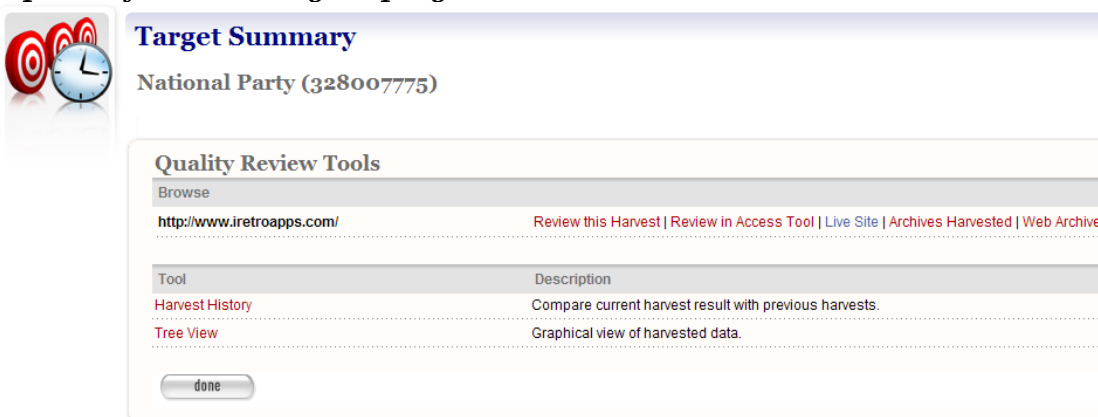


Figure 20. Review Options

- 6** To view the harvested web sites, click the 'Review this Harvest' link below **Browse Tool**.
7 To review the archived web site in the external access tool (Wayback Machine) click 'Review in Access Tool'.
8 To view all archives for the web site in the external access tool click 'Archives Harvested'.
9 To view the site entry page in the public archive (eg: <http://www.webarchive.org.uk>) click 'Web Archive'.
10 To 'prune' the results, click the **Tree View** link.
A nested list of the harvested web sites displays.

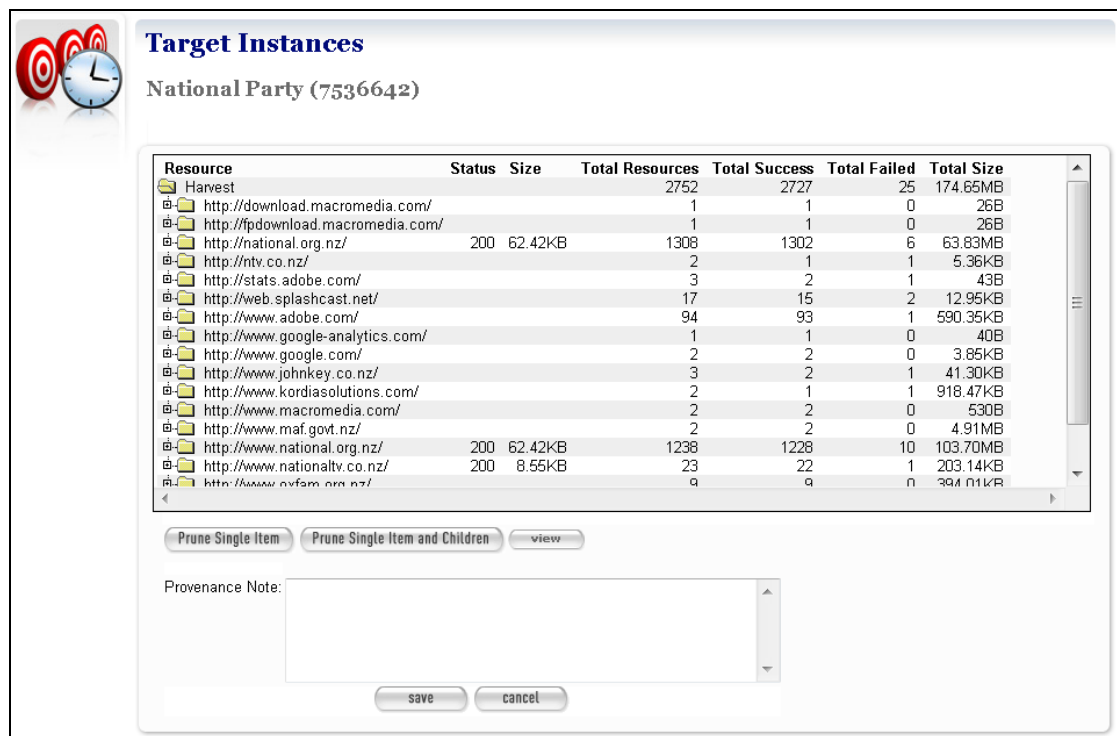


Figure 21. Prune Tool

- 11 To prune the results:
 - click + to expand a hierarchy of sites harvested
 - click to highlight the site you want to prune
 - click **Prune Single Item** to remove just the highlighted page; or **Prune Item and Children** to remove the page and all those listed below it
 - add a note to describe the pruning, and click **Save**.

The display returns to the [Harvest Results tab](#).

Endorse, reject, un-endorse or archive harvest results

- 12 To endorse the results, click **Endorse**.
- 13 To reject the results, click **Reject**.
- 14 To archived an endorsed result, click **Archive**.
- 15 To un-endorse an erroneously endorsed result, click **Un-Endorse**, this sets the target instance back to a harvested state.



Appendix A: Detailed Workflow

