# Contents

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

![Diagram of Web Curator Tool process flow]

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

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

![In Tray screenshot](image)

**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.

![Harvest Authorisations](image)

**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:

![Create/Edit Harvest Authorisations](image)

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.

![Image of Harvest Authorisations](image)

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.

![Image of Harvest Authorisations](image)

Figure 7. Authorising Agencies tab
7 To add a new agency, click **create new**.  
*The Create/Edit Agency page displays.*

![Create/Edit Agency](image)

**Figure 8. Create/Edit Agency**

8 Enter the name, description, and contact information for the agency; and click **Save**.  
*The Authorising Agencies tab redisplays, 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.*

![Permissions](image)

**Figure 9. Permissions tab**

10 To add a new permission, click **create new**.  
*The Create/Edit Permission page displays.*
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 redisplays, showing the added permission.*

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**

Click ✉️ next to the harvest authorisation request.

*The system generates and displays the letter.*
Click to **print** or **e-mail** the letter to the agent.  

*The system sends the letter and changes the permission status to ‘requested’.*

Click **Done**.  

*The [Harvest Authorisations](#) page redisplay.*
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 page screenshot]

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.

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.

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.

![Schedule tab](image)

**Figure 15. Schedule tab**

8 Click **create new**.

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

![Create/Edit Schedule](image)

**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** 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:

![Target Summary](image)

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:

Sortable fields:

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

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

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

**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 instance

**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](image)

**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](image)

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.

1. When viewing a target instance, you can click **to view a target instance, or to edit a target instance**.

   The **View/Edit Target Instance** page displays.
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.*
5 To review a result, click **Review**.

*Options for reviewing display.*

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](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.*
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

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