WEB CURATOR TOOL

Quick Start Guide
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 12
- setting up target schedules for harvesting — see To create a target: page 20
- reviewing harvested content — see To review target instances: page 25.

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.
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:

![Process Flow Diagram]

Figure 1. Web Curator Tool process flow

For a more detailed flowchart, see Appendix A: Detailed Workflow, page 29.
**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 29) — flowchart detailing the complete Web Curator Tool process.
The **Web Curator Tool Home Page** includes the following functions:

![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 harvest 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 — system administrator function to create and manage users, agencies, roles, and privileges.

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](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).

*(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 **Title**, **Authorising Agent**, and/or **Order Number**
- 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.

**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)

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
- **URLs** — 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 **URLs** tab.
The **URLs** tab includes a box for adding URL patterns and a list of added patterns.

**Figure 6. URLs 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.

![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.
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** redisplays, 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.

![Permission Request Letter]

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

The system sends the letter and re-displays the template.

15 Click **Done**.

The **Harvest Authorisations** page redisplays.

**Change the permission status to ‘requested’**

16 Click **edit** next to the harvest authorisation request that includes the permission for which you sent the request letter.

The **General** tab of the **Create/Edit Harvest Authorisations** page displays.

17 Click the **Permissions** tab.

The **Permissions tab** displays.

18 Click **edit** next to the permission for which you sent the request letter.

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

19 Change the **Status** of the permission to ‘requested’, and click **Save**.

20 Click **Save** to close the Harvest Authorisation.
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 page screenshot](image)

**Figure 12. Targets**

At the top of the page are:

- fields to search for existing targets by **Name**, **Seed** (root URL of a Web site), **Agency**, **User**, and **Status**
- 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
- **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

**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 **add**. Repeat for additional sites.

The seed displays in the list below.

You can also use the **Import** button to import a precompiled list of 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.
Click **create new**.

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

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**.

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 Instances](image)

**Figure 17. Target Instances**

At the top of the page are:

- **Fields** to search for existing target instances by **From** and **To** dates, **Agency**, **Owner**, and **Status**.

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

- **View** details
- **Edit** details
- **Delete** a scheduled target instance so its harvest does not run (*target instances can only be deleted in the 'scheduled' state*).
To review target instances:

1. Click 🗿 to view and/or run a target instance, or ✏️ to edit a target instance.

The View/Run/Edit Target Instance page displays.

![Target Instances](image)

Figure 18. View/Run/Edit Target Instance

The View/Run/Edit Target Instance page includes six 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, and archive
- **Annotations** — notes about the target instance.
When viewing (as opposed to editing) a target instance, the **General** tab also includes a **harvest now** button.

2. To run the instance immediately, click **harvest now**.
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 link below **Browse Tool**.

7. To ‘prune’ the results, click **Prune Tool**.

   *A nested list of the harvested web sites displays.*

8. 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, or archive harvest results**

9 To endorse the results, click **Endorse**.
10 To reject the results, click **Reject**.
11 To archived an endorsed result, click **Archive**.
Appendix A: Detailed Workflow

Identify Site to Harvest ➔ Create Permission Record with Assign Approval Task (Pending State) ➔ Approve Permission Task

Authorising Agency Responds ➔ Mail or E-mail Permission Request Letter and Mark Permission as Requested (Requested) ➔ Generate Permission Request

Is Permission Approved? ➔ Mark Permission as Approved (Approved)

No ➔ No

Yes ➔ Mark Permission as Rejected (Rejected)

Create Target (Pending State) ➔ Sets seeds and attach permissions

Review Target ➔ Approve Target Task ➔ Nominate Target (Nominated State)

Is Target acceptable? ➔ Reject Target (Rejected)

No ➔ No

Yes ➔ Tune Target

Approve Target (Approved) ➔ Schedule Instances

Scheduled Time Arrives ➔ Run Harvest ➔ Quality Review Task ➔ Review Target Instance ➔ Is Harvest Okay?

Yes ➔ Is Harvest Okay?

Yes ➔ Can Harvest be fixed?

No ➔ Use other Quality Review tools such as Prune

Endorse Harvest (Endorsed) ➔ Archive Harvest Task ➔ Archive Harvest (Archived)

Can Harvest be fixed?

No ➔ No

Yes ➔ Yes

Legend:

- User Action (Object State)
- System Action
- User Decision
- Task Created in In-Tray
- Time Delay