

Web Harvester Quick Reference

Set Up and Harvest Web Content

Use the Web Harvester from in the Connexion client to associate Web content with a bibliographic record.

	Action
1	<p>Open Connexion client and retrieve the record you want to associate with the Web content you want to harvest.</p> <p>Note: The record must contain an OCLC number (fixed field OCLC) and a title (field 245).</p>
2	<p>On the Tools menu, click Harvest Web Content. The Web Harvester opens in a browser screen.</p>
3	<p>In the Set Up tab, determine the Harvest Parameters Set Up:</p> <p>URL. Enter the URL of the web content you want to harvest.</p> <p>Ignore Robots? Many web sites have a file called <i>robots.txt</i> that may block harvesting.</p> <ul style="list-style-type: none"> Choose Yes to ignore <i>robots.txt</i> when you harvest the site. Choose No to honor <i>robots.txt</i> and not harvest the site. <p>If you ignore <i>robots.txt</i>, OCLC recommends that you inform the site webmaster that you are harvesting the site.</p> <p>Harvest Type.</p> <ul style="list-style-type: none"> Choose By Links to harvest all linked content in the domain, regardless of subdirectory. Choose By Path to harvest only linked content in subdirectories below the entry point. <p>Harvest Depth. For the By Links selection, depth is the number of links away from the entry point. For the By Path selection, depth is the number of directory levels below the entry point.</p> <p>Preview Harvest? Choose Yes to preview the results of your harvest and select the files you want before you complete the harvest. Click No to complete the harvest and then review it.</p> <p>CONTENTdm collection Choose the CONTENTdm collection into which you want to ingest the harvested content.</p> <p>Notes:</p> <ul style="list-style-type: none"> Ingesting content works best with CONTENTdm collections configured with Qualified Dublin Core metadata. The system will pre-populate editable fields with the data entered by the same institution symbol for the previous harvest.
4	<p>Serial Information</p> <p>If the content you want to harvest is a part of a serial, click Yes.</p> <p>Issue Title. Enter the title of the issue.</p> <p>Issue Date. Enter the date of the issue in one of these formats: YYYY-MM-DD or YYYY-MM or YYYY. Note: You must include the dashes (-). (YYYY = 4-digit year; MM = 2-digit month; DD = 2-digit day.) Note: See "Configure CONTENTdm to import issue title and date for serials" below and follow instructions to be sure that your collections accept these fields.</p>

	Action
5	<p>Notification (Optional)</p> <p>Email address. Enter your email address to be notified of the harvest status.</p> <p>Or</p> <p>Check the harvest status later from the Connexion client: On the Tools menu, click Review Web Harvest.</p>
6	<p>When setup is finished, click Harvest to start harvesting, or click Close to cancel.</p>

More about Harvest Depth and Type

When you start harvesting, the Web Harvester uses the **URL** you entered in the **Harvest Parameters Set Up** as an entry point to start gathering content. The Harvester gathers content by following links from the entry point. It does not include content outside the entry point domain (a domain is defined by the part of the URL between the *http://* and the first slash). There are two types of harvesting, Links and Path.

By Links. Use to collect content from anywhere within the domain. The harvester follows all links from the entry point URL. The **Harvest Depth** setting determines how many links away from the entry point to follow.

www.treasury.gov/reports/index.html (entry point)

- ↳ www.treasury.gov/reports/current.html
- ↳ www.treasury.gov/reports/2003/report.pdf
- ↳ www.treasury.gov/reports/2003/report2003.html
- ↳ www.treasury.gov/reports/2003/suppl/s03.pdf
- ↳ www.treasury.gov/reports/2004/agencies.pdf
- ↳ www.treasury.gov/reports/2004/report2004.html
- ↳ www.treasury.gov/funds/funds_in_use.html
- ↳ www.state.gov (not harvested)
- ↳ www.treasury.gov/press_releases.html
- ↳ www.treasury.gov/funds/unclaimed_funds.html

↳ link **DEPTH 0** - Entry point page only **DEPTH 1** - Depth 0 plus all files linked from the entry point **DEPTH 2** - Depth 1 plus all files linked from Depth 1 pages

By Path. Use to limit your harvest based on the directory structure of the Web site. The harvester follows the linked content in subdirectories of the entry point URL. The **Harvest Depth** setting determines how many subdirectories down from the entry point to follow.

www.treasury.gov/reports/index.html (entry point)

- ↳ www.treasury.gov/reports/current.html
- ↳ www.treasury.gov/reports/2003/report.pdf
- ↳ www.treasury.gov/reports/2003/report2003.html
- ↳ www.treasury.gov/reports/2003/suppl/s03.pdf
- ↳ www.treasury.gov/reports/2004/agencies.pdf
- ↳ www.treasury.gov/reports/2004/report2004.html
- ↳ www.treasury.gov/funds/funds_in_use.html
- ↳ www.state.gov
- ↳ www.treasury.gov/press_releases.html
- ↳ www.treasury.gov/funds/unclaimed_funds.html

↳ link **DEPTH 0** - Other linked files in same directory as the entry point **DEPTH 1** - Depth 0 plus all linked files at one subdirectory below the entry point **DEPTH 2** - Depth 1 plus all linked files at one subdirectory level below the Depth 1 files

Not Harvested

Review Harvested Content

To check the status of a harvest or to review harvested content:

	Action
1	In the Connexion client, on the Tools menu, click Review Harvested Content , or if you already have the Web Harvester open, click the Review tab. The list of harvests shows: Date and time when the harvest started; title in the 245 field of the associated bibliographic record; URL of the entry point, OCLC record number, status of the harvest, and a list of actions available for each harvest.
2	If you just harvested content, the list shows the information for the harvest, but the status is In process . Wait a few minutes, and then click Refresh to check that the status changed to Complete .
3	Sort the list. To sort, click a column heading to sort by the data in that column.
4	Actions column. Report. Click to view a report showing the entry point, date, time, status, number of files, total size in megabytes or gigabytes, and the length of time the harvest lasted. Review or Preview. Click to review or preview the harvested content. <ul style="list-style-type: none"> If you selected No for Preview Harvest? in your harvest setup, the Web content is harvested and the list of actions for the harvest contains Review. If you selected Yes for Preview Harvest?, the Web content is retrieved but not yet harvested, and the list of actions contains Preview. <p>See "More about Review versus Preview" below.</p> Ingest. Click to send harvested content to your CONTENTdm collection. If you subscribe to the OCLC Digital Archive, the harvest will also be sent there. Note: Ingesting creates a new 856 field if one was not already in the record. To see the new 856 field, wait a few minutes and retrieve the record again in Connexion client. Delete. Click to delete the harvest. Click OK in the pop-up window to confirm. Note: Clicking Delete removes the harvest only from the list in the Web Harvester. It does not delete harvested content from the Digital Archive or CONTENTdm.

More about Review versus Preview

Review harvested content

In your harvest setup, the default is the Review option (see step 3 in "Set Up and Harvest Web Content" above). When you keep the default and click **Harvest**, content is harvested immediately. To review:

	Action
	In the Action column of the Review list, click Review next to the harvest you want to see. The entry point Web page opens in a separate browser screen.

Preview harvested content

You must select the Preview option in your setup before you harvest content in order to use the Harvest Preview page (see step 3 in "Set Up and Harvest Web Content" above).

With the Preview option selected, when you click **Harvest**, the content is retrieved only for previewing **before** you complete harvest. To preview:

	Action
1	In the Action column of the Review list, next to the harvest you want to see, click Preview . A list of harvested content opens in a separate browser screen.
2	Filter the list. By default the list displays all harvests, with expanded lists of sub-files and/or subfolders, and with all check boxes selected. To change the display, use filters on the left of the screen. View by. Click a button to view Files only or Folders only . Note: If you view Folders only , the other filters are unavailable (grayed out). Click a folder in the Folders only view to display the files within that folder with select/de-select functionality. Show. Click a button to view only selected or deselected files. Limit by File Type. Click a button to view only HTML , Images , PDF , or Word files. Make your selections. <ol style="list-style-type: none"> Click check boxes next to files or folders to select or deselect them for display. Or At the bottom of the filter list, click Select All or Deselect All, and then click check boxes to select or deselect entries. Click a button to view Selected only or Deselected. Note: Your selections/deselections are saved automatically.
2	Important! Whenever you change the list view or use a filter option, you must click Refresh to see the change.
3	Navigate the list. <ul style="list-style-type: none"> Select the maximum number of files to display from Page Size at the top left of the list (default:: 50 records display at a time). Note: The total number of files harvested and the total size of the harvest in megabytes or gigabytes are displayed at the top and bottom left of the list. Display a different part of the list using the arrows at the top and bottom right: <ul style="list-style-type: none"> Use double arrows (<< or >>) to move to the first page or the last page of results. Use single arrows (< or >) to move back one page or forward one page.
4	When your review is completed, click Close to close the Harvest Preview screen. In the Action column of the Web Harvester list, click Harvest to complete the harvest you reviewed, or click Delete to delete it.

Ingest Harvested Content

Action

Once you have reviewed, or previewed and harvested, Web content, in the **Action** list of the Web Harvester Review tab, click **Ingest** to send the harvested content to your CONTENTdm collection and also, if you subscribe, to the Digital Archive.

See more about **Ingest** in step 3 in “Set Up and Harvest Web Content” above.

When the Ingest is complete, the item will appear in the Approve queue in CONTENTdm Administration. You must approve and index the item in order to make it available for search and retrieval via CONTENTdm and the via the link in the WorldCat record (856 field).

Use Reports

Reports describe the results of your harvests. They include the following information:

- **Title.** The title from the 245 field of the record associated with the harvested content.
- **OCLC number.** The OCLC number of the record associated with the harvested content.
- **Harvest date.** The date the harvest was initiated.
- **Status.** Indicates whether the harvest succeeded.
- **Size in megabytes.** Total size of all files in the harvest
- **Files captured.** Total number of files harvested.
- **Harvest duration.** The time the system used to complete the harvest

Configure CONTENTdm to import issue title and date for serials

If you are harvesting issues of a serial and want to include the Issue Title and Issue Date metadata in the item-level metadata record in CONTENTdm, add these metadata fields in your CONTENTdm collection:

- PDI.Title
- PDI.Date

The syntax must be exact:

Enter “PDI” in all caps, followed by a period (.). Enter “Title” and “Date” with initial caps.

If you enter Issue Title and Issue Date in your harvest setup, but have not configured PDI.Title and PDI.Date in your CONTENTdm collection, those metadata elements will not be imported, and your items will not sort appropriately on the interim search results screen.

OCLC Customer Support

OCLC customer service hours: 7:00 am–9:00 pm US Eastern time, Monday through Friday

Telephone:

- USA and Canada: 1-800-848-5800
- International and Central Ohio: 1-614-793-8682

Email: support@oclc.org