Louisiana State Documents
Digital Archive Web Harvesting

LOUIS Users Conference 2009

Sally Wellman
Head of Cataloging
State Library of Louisiana
What you need to get started

- Basic knowledge of cataloging remote electronic resources
- OCLC authorization that permits web harvesting
- OCLC Connexion Client configured for web harvesting
- Patience
What you don’t need

- Access to ContentDM administration
  - LSL will run the indexing every night
  - LSL will make corrections to metadata in ContentDM
Locate a bib record in OCLC or create a new one (if needed)
Prepare to harvest
Gathering metadata
And now you wait…

- Depending on the document, this could take a few seconds or several days.
- You can catalog and harvest multiple items simultaneously.
- If you entered your email in the metadata, OCLC will notify you when the harvesting is complete.
Review harvested content.....
.....or

<table>
<thead>
<tr>
<th>Started</th>
<th>Title</th>
<th>URL</th>
<th>OCLC</th>
<th>Status</th>
<th>Actions</th>
</tr>
</thead>
</table>
Review the links
Ingest if good
Delete if bad (and begin again)
FAQs

- LSL will provide detailed information on local cataloging practices for electronic Louisiana documents.
- LSL is developing a simple electronic “tracking tool” to deliver information about the documents to be harvested.
- LSL will have your authorizations set up for web harvesting.
Web Harvester Quick Reference

Harvest

Harvesting associates web content with a record.

**Action**

1. In Connexion client, retrieve the record you want to associate with the web content you want to harvest.
   **Note:** The record must contain an OCLC number (fixed field OCLC) and a title (field 245).
3. Setup harvest CONTENTdm collection. Choose the CONTENTdm collection into which you want to ingest the harvested content.
   **Note:** Ingesting content works best with CONTENTdm collections configured with Qualified Dublin Core metadata.
   **URL:** Type the URL of the web content you want to harvest.
   **Ignore Robots:** Many websites have a file called robots.txt that blocks the harvesting. Choose Yes to ignore robot.txt and complete the harvest. Choose No to honor robot.txt and not harvest the site. If you ignore robots, OCLC recommends you inform the site webmaster that you are harvesting the site.
   **Harvest Type:**
   - **Path:** Harvests only linked content in subdirectories below

**Harvest Depth and Type**

Start harvesting by choosing an entry point (URL), which is where you want the harvester to start gathering content. The harvester harvests content only by following links. The harvester does not harvest content outside the entry point domain (A domain is what is between the http:// and the first slash in a URL). There are two types of harvesting:

- **Links**
- **Path**

**Links.** From the entry point URL, the harvester follows all links collecting content from anywhere within the domain. You specify how many links away from the entry point the harvester goes by using the depth setting.

```
www.treasury.gov/reports/index.html (entry point)
  www.treasury.gov/reports/current.html
  www.treasury.gov/reports/2003/suppl/s03.pdf
  www.state.gov (not harvested)
  www.treasury.gov/funds/funds_in_use.html
  www.treasury.gov/funds/declared_funds.html
```
Contact

- Diane Brown, Deputy State Librarian
  - 225-342-4923
  - dbrown@slol.lib.la.us

- Sally Wellman, Head of Cataloging
  - 225-342-4939
  - swellman@slol.lib.la.us

- Ferol Foos, Recorder of Documents
  - 225-342-4929
  - ffoos@slol.lib.la.us

- Judy Smith, Head of Louisiana Section
  - 225-342-2791
  - jsmith@slol.lib.la.us