

Web Publication Harvesting Pilot Project White Paper



GPO is pleased to announce the release of a white paper on the results of the recently completed Web Harvesting pilot project to capture official Environmental Protection Agency (EPA) publications in scope of GPO's information dissemination programs.

- [Web Harvesting White Paper](#)
- [Statement of Work, Attachment 1](#)
- [Criteria and Parameters, Attachment 2](#)
- [Blue Angel Technologies Rules, Attachment 3](#)
- [Information International Associates, Inc. Rules, Attachment 4](#)

The white paper reports on the specific context of the results of the pilot, including a summary of analysis done on the work performed, an assessment of lessons learned, and planned future direction and next steps for further development of the harvesting function to be implemented during Release 2 of GPO's Future Digital System (FDsys). The first public release of FDsys is scheduled for November 2008.

As a first step in learning about automated Web publication discovery and harvesting technologies and methodologies, GPO contracted with two private companies on this pilot. We collaborated to develop rules and instructions that would determine whether EPA content discovered was in scope for GPO's dissemination programs. Three separate crawls were conducted on the sites over a six-month period, and harvester rules and instructions were refined and revised between crawls.

Automated publication harvesting was a topic of discussion at the spring 2007 Depository Library Council Meeting (see [session handout](#)).

Sample Publications from GPO's Web Harvesting Pilot

LSCM staff processed a sample of 300 publications harvested during the EPA Pilot Project. The purpose of working through this sample was to determine workflow and staffing implications as well as to estimate the amount of time that would be required to process all the publications acquired during the EPA Pilot Project.

LSCM is testing two mechanisms for making the publications found to be within scope of the FDLP accessible. The majority of publications in the sample are being made accessible through cataloging records in the Catalog of U.S. Government Publications (CGP). Monographs were cataloged using the new brief bibliographic record format, while serials were cataloged following the CONSER abridged standard.

Following the procedures established during the [brief bibliographic records project](#) , the brief records for the monograph publications included in the sample were created directly in the CGP and have not been exported to OCLC. Given the large number of monographs harvested during the EPA Pilot Project, the brief bibliographic records will not be forwarded to the Cataloging Section for enhancement. To allow for an additional searching mechanism, an added entry for the Environmental Protection Agency has been included in each record.

Currently, LSCM assigns PURLs to live content on the publishing agency's Web site. PURLs are only redirect to GPO's archived copy if the live site is no longer available. As part of this project, LSCM is reconsidering this policy. While processing the sample, a portion of the PURLs were directed to the copy of the publication archived on GPO's server rather than the live version.

At the request of the Depository Library Council, LSCM is also trying to determine if there is a mechanism that enables public access to Web harvested content while these publications are in the queue for brief bibliographic records. LSCM has posted a small portion of the sample to GPO Access using a browse table. Publications made accessible through this mechanism will be cataloged in the CGP in the future.

An analysis of the time required to process this sample from the results of the EPA harvesting pilot project is available [here](#) .

To review the sample publications:

- [CGP](#) : Users may conduct a keyword search for the phrase “EPA pilot project” to review the cataloging records.
- [Browse Table](#)

GPO appreciates the input from the 78 respondents who reviewed and submitted comments on the processing of the 300 publications from the results of the EPA Pilot Project. [View a summary of the comments received](#)