Managing Web Harvested Content: Results from the EPA Harvesting Pilot Project

December 21, 2007

In working toward the goal of building a comprehensive collection of digital content within scope of the Federal Depository Library Program (FDLP), the U.S. Government Printing Office (GPO) conducted its first automated publication harvesting pilot project. The goal of the pilot was to test and develop automated and accurate tools and technologies to discover, assess for scope determination, and harvest online publications from the Environmental Protection Agency (EPA).

During the pilot, two vendors, Blue Angel Technologies and Information International Associates, simultaneously but separately crawled the EPA Web site for official EPA publications from March to September 2006. Both vendors used rules and instructions during the crawls that would determine whether the content discovered was in scope of GPO's dissemination programs. Three separate crawls were conducted on the sites during the pilot, and harvester rules and instructions were refined and revised between crawls. At GPO's request, both vendors crawled several EPA databases during the second and third crawl. Blue Angel identified as within scope of the FDLP and harvested 83,229 publications during the final crawl. Information International Associates identified as within scope and harvested 239,478 publications during its final crawl.

Each vendor harvested a number of different types of publications. Using a random sample of 1000 publications (500 from each vendor), LSCM staff reviewed each file to determine the type of publication harvested. The results below are accurate within ±4% for the third crawl.

Type of Publication	Percentage
Monograph	23%
Serial	3%
Database Result	62%
Web Page	9%
Portions of Publications on GPO Access	.08%

LSCM IMPLEMENTATION ASSESSMENT

Following the completion of the pilot, Library Services and Content Management (LSCM) staff began to study the feasibility of classifying and cataloging the monographs and serials acquired through the pilot. Issues LSCM had to consider in relation to this content included:

- LSCM needed to estimate the amount of time and the staffing implications to process and make this content publicly accessible. Both the time involved and the staffing needs will differ for the different types of publications harvested.
- Approximately 14-16% of the results of the final crawl that the vendors identified as within scope are actually not within scope of the FDLP.
- The duplication between the results of the two vendors is unknown as is the amount of duplication with publications already included in the Catalog of U.S. Government Publications (CGP).
- It is estimated that at least 25% of the within scope content represent only a section or a portion of a complete publication. Crawls by both vendors resulted in the harvesting of only portions of a publication, such as a single chapter or an appendix. In other cases, all sections of a publication were harvested but as separate files.
- Approximately 62% of the results from the EPA Pilot Project are results from a database search. While the Pilot demonstrated that the harvesting tools can acquire material from databases, these results are not suitable for cataloging.
- Currently, GPO 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. This policy results in considerable PURL maintenance for LSCM. Publishing agencies do not always advise LSCM when content is taken down or moved. However, publishing agencies prefer the PURL be directed to the live copy on their Web sites as this increases the visibility of their sites.

To begin to address the issues outlined above and to develop a workflow to process all of the publications harvested during the EPA Pilot Project, LSCM processed a sample of 300 publications.

OVERVIEW OF THE SAMPLE OF 300 PUBLICATIONS

LSCM is testing two mechanisms for making the publications found to be within scope of the FDLP accessible using 300 publications, both monographs and serials, from the results of the final crawl. One hundred and fifty publications were identified from each vendor. The majority of publications in the sample are accessible through cataloging records in the CGP (http://catalog.gpo.gov/). Monographs were cataloged using the new brief bibliographic record format, while serials were cataloged using the CONSER standard record format. 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.

LSCM is also using this sample of 300 publications to examine its policy of assigning PURLs. While processing the sample, a portion of the PURLs of the monograph publications were directed to the copy of the publication archived on GPO's server rather than the live version.

Following the completion of the processing of the sample, LSCM will give the depository community the opportunity to review and comment on both mechanisms of access.

PROCESSING OF THE SAMPLE OF 300 PUBLICATIONS

As with all publications, LSCM staff must complete several processing steps prior to making these publications accessible to the public. Automated Web harvested publications require an additional step to ensure that the whole publication was harvested. If the issuing agency placed the publication online in sections, such as separate PDF files for each chapter of a monograph, LSCM must be sure the harvester acquired all the chapters of that monograph. The sections of the publication may be scattered throughout the results from the vendor so staff must expend time to locate all of the sections. To determine completeness, staff compared the harvested copy with the live version of the publication on the EPA's Web site. Staff also examined each publication for indications that portions of the publication were not harvested by reviewing pagination and tables of contents. Only publications that were complete in a single PDF file were used in this sample.

Following the identification of 300 complete publications, LSCM staff reviewed the content of the publication to determine if it was within scope of the FDLP and/or the Cataloging and Indexing Program. Of the 300 publications gathered for the sample, only eight were found not to be within scope. However, this sample, as discussed above, was not picked randomly. While the pilot demonstrated that automated tools can successfully harvest electronic publications, the pilot was less successful in proving that technology can accurately determine scope. Scope determination remains a manual process. GPO's legacy databases and the CGP were then searched to discover if the publications had already been cataloged as an electronic publication or if a tangible version had previously been distributed.

Determination	Number of Publications	Percentage of the Sample
Already Cataloged as an electronic title	57	18.5%
Previously distributed in tangible format	10	3%
Not within scope	8	2%
New publication	232	62%

A note in the 590 field that reads "An additional copy of this publication was harvested as part of the EPA Pilot Project" was added to cataloging records of all the publications in the sample that had already been cataloged as an electronic publication. A PURL and a note in the 590 field that reads "An additional copy of this publication was harvested as part of the EPA Pilot Project" was added to the cataloging records of all publications in the sample that had previously been distributed in a tangible format. All new serial publications included in the sample were passed to the Bibliographic Control Section so CONSER standard records could be created. Each of these records notes which issues were harvested as part of the EPA Pilot Project in the 590 field.

The remaining monograph publications were divided into two groups. A browse table indicating title, date, and series numbering was created on *GPO Access* (link) using 100 publications from the sample. The remaining publications were added to the CGP following the procedures established during the brief bibliographic records project (http://www.access.gpo.gov/su_docs/fdlp/cip/creation-brief-bib-records.pdf). The brief records for these publications were created directly in the CGP and were not exported to OCLC. To allow for an additional searching mechanism, an added entry for the Environmental Protection Agency was included in each record. Each record contains a note in the 590 field that reads "EPA pilot project". Given the large number of monographs harvested during the EPA Pilot Project, the brief bibliographic records will not be forwarded to the Bibliographic Control Section for enhancement.

PROCESSING TIMES FOR THE SAMPLE OF 300 PUBLICATIONS

Processing Step	Average Time
Identification of a Complete Publication	2 minutes
Scope Determination and Search for Duplicates	17 minutes
Creation of Brief Bibliographic Record	30 minutes (including SuDoc class creation)
Creation of CONSER Standard Record	2 hours 30 minutes (including SuDoc class creation, name authority work, and contact with issuing agency if needed)
Add PURL to Publications Distributed in Tangible Format	7.5 minutes
Creation of Browse Table	4 hours total to create the entire browse table (including renaming files)

ESTIMATED PROCESSING TIMES FOR ALL MONOGRAPHS AND SERIALS

Monographs

An estimated 74,222 monographs were harvested by both vendors during the final crawl of the EPA's Web sites. The numbers below assume that 14% will not be within scope, that 18.5% of the publications will already have a cataloging record in the CGP, and that 3% will have previously been distributed in a tangible format.

Processing Step	Average Time
Identification of a Complete Publication	2,474 hours for publications complete in a single file
Scope Determination and Search for Duplicates	21,029 hours
Creation of Brief Bibliographic Record	25,054 hours (including SuDoc class creation)

Add PURL to Publications Distributed in Tangible Format	239 hours
Total:	48,796 hours

Serials

An estimated 9,681 serial issues* were harvested by both vendors during the final crawl of the EPA's Web site. The numbers below assume that 14% will not be within scope, that 18.5% of the publications will already have a cataloging record in the CGP, and that 3% will have previously been distributed in a tangible format.

Processing Step	**Average Time**
Identification of a Complete Publication	323 hours for publications complete in a single file
Scope Determination and Search for Duplicates	2,743 hours
Creation of CONSER Standard Record	16,340 hours* (including SuDoc class creation, name authority work, and contact with issuing agency if needed)
Create Separate CONSER Standard Record for Electronic Version of Titles Distributed in Tangible Format	725 hours*
Total:	20,131 hours*

*LSCM was unable to estimate the number of serial titles versus the number of serial issues so the time required to create the needed CONSER records may be lower.

CONCLUSION

LSCM believes that providing access to the monographs and serials harvested as part of the EPA Pilot Project via the CGP best serves the needs of the depository community and the general public. As can be seen from the sample of 300 publications, making the content from the EPA Pilot Project accessible to the public is a multi-step process and involves the commitment of a significant amount of time. However, as staff become more familiar with the new brief bibliographic record format the time required to create one of these records will decrease. The identification of complete publications, the identification all the parts or issues of a title scattered within the results of the harvest and the de-duplication of the contents will continue to require a significant amount of time and staff to complete.

Additionally 1,000 monographs within scope of the FDLP have been identified from EPA Pilot Project for inclusion in the Automated Metadata Extraction Project. This is a two year project with the Defense Technical Information Service (DTIC) and Old Dominion University (ODU) to use automated metadata extraction software tools to create metadata for groups of electronic publications in GPO's electronic collection. This is a two year project and the results are not expected until near the end of the project.

After reviewing the comments of the depository community on the two mechanisms of access, LSCM will begin processing the remaining content from the EPA Pilot Project. The output of this effort for content within scope will be brief bibliographic records for the monographs and CONSER standard records for the serials. Further analysis is needed of the non-traditional content, such as the databases results, before any processing work can be completed on that content. As the work on processing this material begins, LSCM asks that the depository community keep in mind that staff must continue to acquire and catalog new publications, both electronic and tangible, from sources other than the EPA harvest results.