GPO Web Harvesting Pilots and Future Initiatives

U.S. GOVERNMENT PRINTING OFFICE
KEEPING AMERICA INFORMED

October 23, 2006

Introduction

Over the past few years, many in-scope digital publications being issued by Federal agencies are not being included in the Federal Depository Library Program (FDLP) or the Cataloging and Indexing Program. GPO is frequently not informed of these new publications, known as “fugitive publications”, when they are published directly to the Web and not sent through GPO printing procurement processes.

In working toward the goal of building a comprehensive collection of content available through its dissemination programs, GPO will implement a set of automated tools and technologies that can identify and harvest fugitive publications from agency Web sites.

As a first step in learning about technologies and methodologies, GPO recently completed a pilot project with two private companies that harvested in-scope publications from the Environmental Protection Agency’s Web sites. Three separate crawls were conducted on the sites over a six-month period, and harvester rules and instructions were refined and revised between crawls. GPO will leverage the knowledge it acquires to build a set of requirements for the comprehensive harvesting solution to be implemented in conjunction with the Future Digital System (FDSys).

General Assumptions for GPO Long-Term Harvesting

1. GPO will use discovery, assessment, and harvesting tool(s) to identify, gather, and capture official publications from Federal Agency Web sites.

2. The harvesting function will be performed either by GPO internally, an outside contractor, or a combination of the two.

3. Federal Agencies will expect GPO to notify them that it is harvesting publications from their Web sites, even though the information is posted for public access.

4. The harvester (including discovery, assessment, and harvesting tools) will be fully implemented in conjunction with GPO’s Future Digital System (FDSys).

5. The harvesting function will retrieve content and metadata necessary to create a package for ingest into the FDSys, but additional processing will be required in order to complete the package for ingest into FDSys.

6. Harvesting activities will follow industry best practices to ensure that GPO and target servers are not put at risk in terms of security and bandwidth.

Pilot process

The pilots were conducted by each contractor separately using different technologies and methodologies. Each contractor performed the following tasks:

- Crawled the EPA Web sites and any linked outside domains to identify in-scope publications.
- Using criteria and parameters provided by GPO, applied rules to determine if a publication is within scope of GPO's dissemination programs.
- Harvested any available content and metadata if content was determined to be in scope.
- Refined rules that determine scope between each crawl based on GPO analysis of the results.
- Compared the results of publications deemed to be in-scope with all EPA MARC records currently in the Catalog of U.S. Government Publications (CGP) to determine what online publications have not yet been brought under bibliographic control.
- Wrote rules for and harvested content from selected EPA query-based databases.

Preliminary Pilot Results

- Contractor #1 discovered and harvested 83,229 documents* that their technologies and rules have deemed to be in scope of the FDLP.

- Contractor #2 discovered and harvested 239,478 documents* that their technologies and rules have deemed to be in scope of the FDLP.

- The accuracy rate of scope determination for the pilot is projected to be between 70% and 85% based on initial GPO sampling and analysis.

* "Documents" refers to complete publications or individual parts of publications.

Lessons Learned

- While many rules used to determine scope can be aggregated to other agencies, GPO believes there will be a certain amount of customization required for each target Web site in order to achieve maximum accuracy and comprehensiveness, as well as ensure that intended use and access rights are followed for harvested content.

- The assessment of whether content is within the scope of the FDLP has historically been a largely subjective process that is based on the experience gained by GPO personnel. As expected, it has been difficult to mimic this traditionally subjective decision with objective rules, based on the projected error rates reported above.

- Publications in certain file formats, including PDF, MS Office applications, and text, were more easily harvested in their entirety than those in HTML or other file formats.

- Publications that are comprised of multiple files proved to be a challenge in that it was difficult to write rules that related the various pieces of a publication together.

Next Steps

- The results of the pilot are currently being evaluated by GPO. A white paper on the final results of the pilot will be published by GPO in November.

- GPO will continue to review and compare results of its pilots with similar projects (e.g., NDIIPP Initiatives).

- Based on lessons learned above and issues that still need to be resolved, more information is needed to take place before the harvester is implemented in conjunction with FDsys. To this end, GPO plans to conduct another pilot with other agency Web sites that will test similar technologies and methodologies based on lessons learned.

- The knowledge gained from the pilots will be leveraged in the implementation of the Harvester (including discovery, assessment, and harvesting tools) as a part of FDsys.

- GPO has received all digital content from the pilot. Our goal is to catalog in-scope publications harvested from this pilot, following an investigation of the results of the automated comparison between the third crawl results with EPA records from the Catalog of U.S. Government Publications.

- While automated publication harvesting technology solutions are investigated and developed to improve accuracy and comprehensiveness, GPO will continue to identify and harvest publications.

Discussion Questions

1. Are the assumptions stated above correct with respect to Web Harvesting?

2. A harvester can be configured to harvest *only* in-scope publications or *mostly* in-scope publications including some out-of-scope publications. Given the results of the pilot, is the existing methodology sufficient to continue harvesting?

3. What other avenues regarding automated Web Harvesting should GPO be exploring in the future?

4. Do you have suggestions for which agency Web site(s) would be best to focus on in a future pilot?