# Managing Web Harvested Content

U.S. GOVERNMENT PRINTING OFFICE
KEEPING AMERICA INFORMED

April 17, 2007

## Introduction

One mission of the U.S. Government Printing Office (GPO) is to provide permanent public access to U.S. Government information products. A primary goal of GPO's *Strategic Vision for the 21st Century* is the implementation of a digital repository for all Federal publications – past, present, and future. One way to acquire current publications for GPO's information dissemination programs is to capture, or harvest, online U.S. Government publications.

GPO recently completed 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).

GPO's ongoing manual and semi-manual harvesting efforts are also adding content to the comprehensive collection. Manual harvesting identifies and captures known online publications and their associated files. Semi-manual harvesting uses a software tool to schedule content capture and reharvest content at known Web sites.

Library Services and Content Management (LSCM) is now in the process of developing an overall plan to manage the acquisition, classification, cataloging, and storage of all Web harvested content, including publications acquired through the pilot.

## General Assumptions for Managing Web Harvested Content

- GPO will continue to develop more fully automated publication harvesting tools and methodologies as part of the Future Digital System (FDsys).
- GPO will continue to manually and semi-manually harvest known publications.
- GPO will make harvesting decisions in accordance with the LSCM collection development policy which is currently under development.
- LSCM will continue to be responsible for scope determinations of harvested content and for classification and cataloging of publications deemed to be in-scope for GPO's information dissemination programs, even after implementation of the FDsys.
- Material inadvertently harvested but not within scope of the Federal Depository Library Program (FDLP) or the Cataloging and Indexing Program (C&I) will not be retained by GPO.
- LSCM may use a combination of in-house cataloging, cataloging contracts, and automated metadata extraction to create bibliographic records for Web harvested publications.
- Bibliographic records for Web harvested publications will be completed in accordance with overall cataloging priorities.
- Bibliographic records for Web harvested publications will be created at an abridged or other brief level. The depository library community will be given the opportunity to review new standards prior to implementation. As cataloging practices change over time, the GPO cataloging standards for Web harvested publications may change as well.
- LSCM will explore the use of automated metadata extraction tools as a method to create bibliographic records for Web harvested publications.

## Issues Related to Managing Web Harvested Content

The management of Web harvested content touches on a number of issues that affect both policies and operations within LSCM.

Assignment of Persistent Uniform Resource Locators (PURLs) or successor system:

- Currently, GPO assigns PURLs to live content on the publishing agency's Web site. PURLs are redirected to GPO's archived copy only if the live site is no longer available.
- Publishing agencies prefer the PURL be directed to the live copy on their Web sites. This increases the visibility of their Web sites.
- The current policy results in considerable PURL maintenance for LSCM. Publishing agencies do not always advise LSCM when content is taken down or moved.

Superintendent of Documents Policies

- The need to manage a large and growing amount of Web harvested content necessitates the review of a number of policies and the possible development of new ones.
- These policies may include harvesting, collection development, cataloging priorities, and scope determination.

Cooperative Cataloging

- LSCM is interested in exploring the use of cooperative cataloging partnerships as an additional method for completing bibliographic records for Web harvested content.
- Before partnerships of this type could be established, procedures and quality control mechanisms must be in place.
- LSCM must also complete the testing of its Z39.50 gateway to allow for easy transfer of bibliographic records.

Harvesting Complete Publications

- Reviews of EPA pilot project results revealed that 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.
- It is estimated that at least 25 percent of the within-scope content represents only a section or a portion of a complete publication.
- While more analysis is needed, GPO anticipates that reconstruction of these partial publications and elimination of duplication will be very time consuming.

## Ongoing Technology Discovery

GPO will continue to work to develop more fully automated publication harvesting tools and methodologies in preparation for full implementation under the FDsys. While the pilot demonstrated that scope determination of online documents discovered and harvested can be automated to a reasonable extent, much additional processing is needed in order for harvested content to be made available via FDsys and the FDLP. These processes include:

- Grouping of portions of documents into entire publications.
- Inspection of harvested content: review of content harvested for accuracy of scope determination.
- Cataloging and classification: creation of cataloging records and classification for in scope content.

As part of its continuing technology discovery efforts, GPO plans to explore automated ways of performing the functions above. This may be accomplished with tools that are separate from harvesting technologies.

## Questions for Discussion

1. Are the assumptions stated above correct with respect to processing Web harvested publications?

2. Should Web harvested publications be identified as such in the Catalog of U.S. Government Publications?

3. Should LSCM point PURLs at the live copy of a publication on the agency Web site or at the archived copy on a GPO server?

4. Are cooperative cataloging partnerships an avenue LSCM should explore to assist with the creation of bibliographic records for Web harvested publications?

5. Are the cataloging levels outlined above acceptable?

6. Are there groups of publications that should be among those manually or semi-manually harvested by LSCM?

7. What should GPO do with out-of-scope material accidentally harvested?

U.S. GOVERNMENT PRINTING OFFICE I KEEPING AMERICA INFORMED

# Managing Web Harvested Content

April 17, 2007

# Introduction

The U.S. Government Printing Office (GPO) is developing a digital repository for all Federal publications – past, present, and future.

One way to acquire current publications for GPO's information dissemination programs is to capture, or harvest, online U.S. Government publications.

# Harvesting Pilot Project

GPO recently completed 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).

# Ongoing Harvesting Activities

GPO's ongoing manual and semi-manual harvesting efforts are also adding content to the comprehensive collection.

- Manual harvesting identifies and captures known online publications and their associated files.
- Semi-manual harvesting uses a software tool to schedule content capture and reharvest content at known Web sites.

# Harvested Content Management Plan

Library Services and Content Management (LSCM) is now in the process of developing an overall plan to manage the acquisition, classification, cataloging, and storage of all Web harvested content, including publications acquired through the pilot.

# Major Issues

- Assignment of Persistent Uniform Resource Locators (PURLs) or successor system
- Superintendent of Documents Policies
- Cooperative Cataloging
- Harvesting Complete Publications

# Persistent Uniform Resource Locators

- Currently, GPO assigns PURLs to live content on the publishing agency's Web site. PURLs are redirected to GPO's archived copy only if the live site is no longer available.
- Publishing agencies prefer the PURL be directed to the live copy on their Web sites. This increases the visibility of their Web sites.
- The current policy results in considerable PURL maintenance for LSCM.

# Superintendent of Documents Policies

- The need to manage a large and growing amount of Web harvested content necessitates the review of a number of policies and the development of new ones:
  - SOD 304: Harvesting Federal Digital Publications for GPO's Information Dissemination Programs
  - Defining in-scope publications
  - Collection development
  - Cataloging priorities

# Cooperative Cataloging

- LSCM is exploring the use of cooperative cataloging partnerships as an additional method for completing bibliographic records for Web harvested content. Procedures and quality control mechanisms must be in place for these partnerships.
- LSCM must also complete the testing of its Z39.50 gateway to allow for easy transfer of bibliographic records.

# Harvesting Complete Publications

- Reviews of EPA pilot project results revealed that 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.
- It is estimated that at least 25 percent of the within-scope content represents only a section or a portion of a complete publication.

# Ongoing Technology Discovery

GPO will continue to work to develop more fully automated publication harvesting tools and methodologies in preparation for full implementation under the FDsys.

Issues related to automating this process include:

- Grouping of portions of documents into entire publications.
- Inspection of harvested content: review of content harvested for accuracy of scope determination.
- Cataloging and classification: creation of cataloging records and classification for in scope content.

# SOD 304 Policy Statement

"GPO will acquire publications for inclusion in the National Bibliography and the FDLP through manual and automated harvesting. GPO will use automated harvesting programs only with the publishing agency's advice and prior consent. Permission to manually harvest publications from agency publicly accessible Web sites will not be sought."

http://www.access.gpo.gov/su_docs/fdlp/pubs/policies/sod_304(final).pdf

# SOD 304 Policy Statement Review

This internal policy provides guidance and instruction for harvesting of publications from Federal agency Web sites

Changes may be made in SOD 304 as related policies are reviewed and developed that affect the management of the harvesting process as well as the files themselves

# Brainstorming

- Related Policies and Procedures
  - Scope
    - FDLP (section 1902)
    - Cataloging and Indexing (sections 1710-1711)
    - Online Access to Publications (section 4101)
    - Publishing Agency guidance
      - OMB Circular A-130
      - E-Government initiatives
  - Cataloging Priorities (in Cataloging Guidelines)
  - Collection Development

# General Assumptions

GPO will continue to develop more fully automated publication harvesting tools and methodologies as part of the Future Digital System (FDsys).

GPO will continue to manually and semi-manually harvest known publications.

GPO will make harvesting decisions in accordance with the LSCM collection development policy

# General Assumptions

LSCM will continue to be responsible for scope determinations of harvested content and for classification and cataloging of publications deemed to be in-scope for GPO's information dissemination programs, even after implementation of the FDsys.

Material harvested but not within scope of the Federal Depository Library Program (FDLP) or the Cataloging and Indexing Program (C&I) will not be retained by GPO.

# General Assumptions

LSCM may use a combination of in-house cataloging, cataloging contracts, and automated metadata extraction to create bibliographic records for Web harvested publications.

Bibliographic records for Web harvested publications will be completed in accordance with overall cataloging priorities.

# General Assumptions

Bibliographic records for Web harvested publications will be created at an abridged or other brief level. The depository library community will be given the opportunity to review new standards prior to implementation. As cataloging practices change over time, the GPO cataloging standards for Web harvested publications may change as well.

LSCM will explore the use of automated metadata extraction tools as a method to create bibliographic records for Web harvested publications.

# Questions for Discussion

1. Are the assumptions stated above correct with respect to processing Web harvested publications?

# Questions for Discussion

2. Should Web harvested publications be identified as such in the Catalog of U.S. Government Publications?

# Questions for Discussion

3. Should LSCM point PURLs at the live copy of a publication on the agency Web site or at the archived copy on a GPO server?

# Questions for Discussion

4. Are cooperative cataloging partnerships an avenue LSCM should explore to assist with the creation of bibliographic records for Web harvested publications?

# Questions for Discussion

5. Are the cataloging levels outlined above acceptable?

# Questions for Discussion

6. Are there groups of publications that should be among those manually or semi-manually harvested by LSCM?

# Questions for Discussion

7. What should GPO do with out-of-scope material accidentally harvested?