# Statement of Work (SOW) for Web Harvesting
# U.S. Government Printing Office
# Office of Information Dissemination

## Scope

The U.S. Government Printing Office (GPO) requires the services of a vendor that can provide a number of different products and/or services related to the discovery, harvesting, and assessment of documents and publications from Web sites using Web crawler and other appropriate technologies (to be specified by vendor). GPO is involved in a project that is attempting to discover and retrieve publications from Federal agency Web sites in order to identify publications that have not been cataloged by GPO but fall within the scope of the Federal Depository Library Program (FDLP) and the National Bibliography.

## Background on the FDLP

The FDLP was established by Congress to ensure that the American public has access to its Government's information. Since 1813, depository libraries have safeguarded the public's right to know by collecting, organizing, maintaining, preserving, and assisting users with information from the Federal Government. The FDLP provides Government information at no cost to nearly 1,250 depository libraries throughout the country and territories. These depository libraries, in turn, provide local, no-fee access to Government information in an impartial environment with professional assistance.

GPO manages the National Bibliography Program and is responsible for maintaining Franklin (formerly known as the *Catalog of United States Government Publications*). Franklin is comprised of bibliographic records of U.S. Government information products published by all three branches of the U.S. Government that are included in the FDLP. Bibliographic records are added daily to Franklin, with approximately 22,000 records added annually. Franklin links users directly from bibliographic citations to electronic publications by using PURLs (Persistent Uniform Resources Locators) or by assisting the public in locating information in depository libraries and through the GPO Sales Program. GPO bibliographic data is also available to individual libraries directly from GPO and from a variety of commercial sources. This data can be used to populate local databases and public access catalogs with bibliographic citations for U.S. Government publications.

GPO prepares machine-readable cataloging records (MARC) for the Online Computer Library Center (OCLC) bibliographic network. Library Technical Information Services within the Office of Information Dissemination at GPO is the national authority for cataloging and bibliographic control of U.S. Government information products and is an active partner in all components of the Library of Congress' Program for Cooperative Cataloging. In addition, GPO prepares and adheres to the *GPO Cataloging Guidelines*,

which provide specific guidance for cataloging complex and dynamic U.S. Government publications and are an essential resource for the National Bibliography Program.

## Background on the GPO's Future Digital System

GPO is working to develop GPO's Future Digital Information System. As outlined in the Strategic Vision, this Digital Content System will allow federal content creators to easily create and submit content that can then be preserved, authenticated, managed and delivered upon request. This Future Digital System (FDsys) will form the core of GPO's future operations.

Included in the FDsys will be all known Federal Government documents within the scope of GPO's Federal Depository Library Program (FDLP), whether printed or born digital. This content will be entered into the system and then authenticated and catalogued according to GPO metadata and document creation standards. Content may include text and associated graphics, video and sound and other forms of content that emerge. Content will be available for Web searching and Internet viewing, downloading and printing, and as document masters for conventional and on-demand printing, or other dissemination methods.

GPO has identified three main types of content that the system will be managing:

- Deposited content: Content intentionally submitted to GPO by Content Originators (e.g. Federal agency Publishers).
- Converted content: Digital content created from a tangible product (e.g., scanned digital documents).
- Harvested content: Content within the scope of GPO dissemination programs that is gathered from Federal agency Web sites.

The focus of this SOW will be on harvested content, specifically pointing towards the development of a "Harvester," which will include Discovery, Assessment, and Harvesting Tools that will be used to harvest content to be included in the FDsys. Discovery tools will locate electronic content from Federal agency Web sites and provide information to the assessment tool. Assessment tools will determine if the discovered content is within the scope of GPO dissemination programs and whether other versions of the content already exist in the system and establishes appropriate relationships between versions. Harvesting tools gather content and available metadata.

For more information on the FDsys, including the Concept of Operations and Requirements Documents, please go to: http://www.gpo.gov/projects/fdsys.htm.

## GPO's Web Harvesting Project

Over the past few years, GPO has become increasingly aware that many publications being published by Federal agencies are not being included in the FDLP; these documents have come to be known as "fugitive publications". With increasing frequency, agencies are publishing content only in electronic formats and, when this occurs, they frequently fail to inform GPO of these new publications for inclusion in the FDLP and Franklin. In addition, agencies sometimes procure their printing directly from private sector companies or use in-house facilities rather than coming to GPO and then fail to inform GPO of these publications, although there may be electronic counterparts on the publishing agency Web sites that could and should be included in the FDLP and Franklin.

In light of the large number of publications that have become fugitive, GPO is seeking Web crawler and other technologies that can provide a solution for the identification and harvesting of fugitive documents and publications from agency Web sites. In order to begin, GPO plans to launch a pilot project with the Environmental Protection Agency (EPA) to crawl the primary EPA Web site and its sub-agency Web sites.

This project will be instrumental in the formation of long term requirements and specifications for portions of the FDsys. GPO plans to leverage what it has learned in this pilot to build a comprehensive harvesting solution in conjunction with the implementation of the FDsys.

NOTE: GPO is seeking contractors that CURRENTLY possess the capabilities and technologies to perform the tasks below. It is not the intention of GPO to contract with a vendor that is planning to build these technologies during its relationship with GPO.

**Overall Goal for Harvesting and Objectives for this SOW**

Overall Goal for Location and Harvesting: To discover, identify, and harvest electronic publications residing on Federal Agency Web sites (starting as a pilot with the Environmental Protection Agency) that that have not previously been a part of GPO's electronic collection but fall within the scope of the Federal Depository Library Program (FDLP) and National Bibliography.

Objectives for this SOW in Support of Locating and Harvesting:

1. To identify, learn about, utilize web crawling and other applicable technologies (to be specified by the contractor) that can discover, assess, and harvest electronic Government Publications on Federal Agency Web sites based on a flexible set of rules and instructions that are derived from criteria being developed by GPO on the characteristics of publications that fall within the scope of the FDLP.
2. To identify, learn about, and utilize a tool that can accurately provide automated comparison and collections analysis, in order to determine whether the harvested documents have already been cataloged by GPO in electronic format. The tool will weigh the listing of publications harvested from the Web crawler against the

listing of tangible and electronic EPA publications that have already been cataloged by GPO in the FDLP or that are retrieved from prior crawls of the selected websites.

3. To assess the accuracy by which the technology can identify electronic publications that fall within the scope of the FDLP, and to leverage the knowledge acquired from this pilot to further develop the requirements and specifications for the implementation of Discovery, Assessment, and Harvesting Tools in conjunction with the FDsys.

**Metrics and Benchmarking**

The benchmark and metrics will be used to evaluate the level of success achieved during this project.

Benchmark #1: The results of the three crawls being performed will be assessed (through the metrics below) based on the manual harvest that is currently being conducted by GPO staff of the EPA Web site.

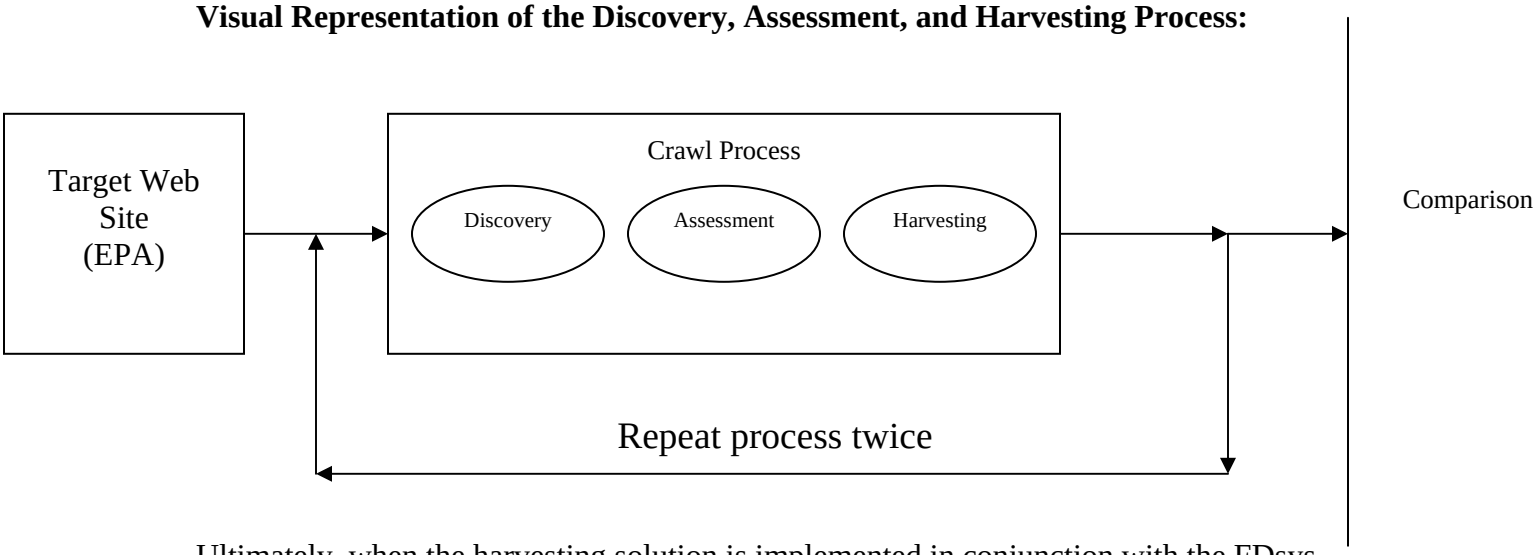
- *Metric:* Accuracy of publications located by the crawler
    - By the end of the third crawl, a maximum of 10% error between the **number** of in-scope documents harvested by the crawler technology and that of benchmark #1 (manual harvest conducted by GPO staff).

Benchmark #2: The documents located by the crawling technology will be evaluated based on a manual process of determining whether the documents harvested fall within the scope of the FDLP (NOTE: due to the large number of documents that are anticipated to be harvested, this manual process may initially be applied to a representative fraction of the documents harvested for the purposes of assessing the project).

- *Metric:* Accuracy of located documents judged to be within scope
    - By the end of the third crawl, a maximum of 1% error of in-scope documents harvested by the crawler technology based on manual assessment by GPO staff (i.e. a 99% similarity between the publications harvested by the crawler technology versus benchmark #2).

NOTE: The metrics listed above are guidelines for measuring the success of this pilot project. It is expected that a certain amount of improvement will be seen in the second and third crawls, given that the set of rules and instructions used by the crawler will be modified based on the results of previous crawls. The metrics above are not absolute measures of success or failure of the project, but instead are best estimates of guidelines for the success of the Web crawling technology.

**Visual Representation of the Discovery, Assessment, and Harvesting Process:**




Ultimately, when the harvesting solution is implemented in conjunction with the FDsys, the end result of the process above will be the creation of Harvested Content Packages containing all content and corresponding metadata necessary to create a Submission Information Package (SIP) that complies with content standards for the FDsys.

## Contractor Tasks

The key capabilities GPO is seeking in relation to this project are to provide Web crawling and other technologies (contractor specified) that will locate, identify and capture all publications from all pages on the EPA Web site and its sub-agency Web sites that fall within the scope of the FDLP. A preliminary set of tasks is mapped out below.

1. Based on criteria currently being developed by GPO for the characteristics that constitute a publication, build a set of rules and instructions for the crawler technology to capture all documents that meet these criteria. This must include the capability to refine and revise rules and instructions over time, as GPO gets further along the learning curve.
    - Rules and instructions should be developed in collaboration with all relevant areas of GPO, including but not limited to the Program Management Office, the Office of Information Dissemination and the Office of the CIO.

2. Work with GPO personnel to set up the parameters for the crawl of the EPA Web site, in order to ensure that all relevant areas of the EPA and its sub-agencies are being crawled.

3. Conduct the first crawl of the EPA Website and build a list of publications available on the Web site.
    a. Identify publications in all possible formats, such as HTML, PDF, MS Word and Excel files, etc.

b. Crawl and harvest the content of each publication, as well as external and internal metadata tied to each file. Required metadata includes, but is not limited to:
    i. Descriptive (e.g., title, date)
    ii. Structural (e.g., parent/child relationships)
    iii. Technical (e.g., file format, MIME type)
    iv. Administrative (e.g., rights information, creator/originator)

    NOTE: Please see the FDsys System Requirements Document (pages 32-34) for more detail on metadata requirements for the FDsys, located at: <http://www.gpo.gov/projects/pdfs/FDsys_RD_v1.0.pdf>

c. Perform automated elimination of those publications retrieved by the crawler that do not fall within the scope of the FDLP and National Bibliography based on GPO's set of criteria.
d. Identify and report all versions/editions of publications that may have multiple versions or additions.

    NOTE: Harvesting in scope documents from the surface pages of the EPA websites is the minimum requirement. However, if applicable, contractors should also provide an explanation of a solution that discovers and retrieves in-scope documents from the "hidden web" (e.g., content that resides in query-based databases or Agency Content Management Systems) in their proposals.

4. Using data collected from manual "crawling" conducted by GPO and in conjunction with GPO personnel, further refine the parameters for the next crawl of the EPA Web site.

5. Conduct the second crawl of the EPA Web site using the newly refined parameters set forth during task 4, performing once again duties a, b, c, and d that were performed under task 3.

6. Using data collected from manual "crawling" conducted by GPO, further refine the parameters for the next crawl of the EPA Web site.

7. Conduct the third crawl of the EPA Website using the newly refined parameters set forth during task 6, performing once again duties a, b, c, and d that were performed under task 3.

8. Conduct automated comparison/collections analysis. Publications retrieved from the Web crawler and other technologies will be matched against one or more publication databases provided by GPO, one of which will be based on MARC records cataloged for the FDLP and Franklin.
    a. Retain information in a database about all items harvested in order to avoid duplications in subsequent crawls.

b. Match the publications harvested with those already cataloged by GPO, using not only the Web site file location, name, size and date, but all relevant content and metadata as well.
c. Identify publications not already harvested, but already cataloged by GPO based on print or microfiche editions in the FDLP, and subsequently associate the harvested electronic file with that record.
d. Identify publications that have not been harvested or cataloged by GPO.

## Deliverables, Products

For Deliverable Products #1-7, the contractor shall furnish 1 hard copy and send electronic copies of the reports to designated GPO contacts (to be determined).

NOTE: Any business rules created by the contractor as a work product of this contract relating to Web harvesting and/or collections analysis will become the sole property of the Government Printing Office. The contractor shall deliver to GPO:

Deliverable Product # 1: A report clearly presenting in its text the set of rules and instructions developed for the crawler technology to capture only those documents that meet the criteria. These instructions should be based on criteria developed by GPO for the characteristics that constitute a publication. The report should state that these rules could be modified or changed over time and explain in detail what time and resources would be required to do so. Information for this report is derived from Contractor Task 1.

Deliverable Product # 2: A report to GPO outlining the results of the first crawl of the EPA Web site. The report should first outline all background information on the crawl, including: procedures followed, timeframes for the duration of the crawl, any issues or obstacles observed, and any other relevant background information. The report should then provide a comprehensive listing of all publications retrieved during the harvest, stating explicitly the titles and file formats of each, as well as the amount of information crawled for each (i.e. what content and/or internal and external metadata was retrieved). The report should also provide a listing of publications crawled that do not fall within the scope of the FDLP based on criteria set forth by GPO, and also a separate listing of those publications that have multiple versions/editions. Information for this report is derived from Contractor Tasks 2 and 3.

Deliverable Product #3: A report clearly presenting in its text the refined set of rules and instructions developed for the crawler technology to capture all documents that meet the criteria. These instructions reflect the further refinements to the set of rules and instructions resulting from the completion of Contractor Task 4.

Deliverable Product #4: A report to GPO outlining the results of the second crawl of the EPA Web site. This report should be in the same format as deliverable product #2, but should mainly focus on the improvements made since the last crawl based on the refinement of rules and instructions. Along with the comprehensive listing of all publications retrieved during the harvest, it should separate out the new publications

retrieved and provide insight into what change in the rules and instructions allowed for the harvest of these new documents. The second crawl should NOT exclude in-scope documents retrieved in the previous crawl. Information for this report is derived from Contractor Task 5.

Deliverable Product # 5: Repeat of Deliverable Product #3. Information for this report is derived from Contractor Task 6.

Deliverable Product # 6: Repeat of Deliverable Product #4. Information for this report is derived from Contractor Task 7.

Deliverable Product # 7: A report to GPO summarizing the automated comparison/collections analysis conducted by the contractor. The report should first outline all background information on the analysis, including: procedures followed, timeframes for the duration of the analysis, any issues or obstacles observed, and any other relevant background information. The report should then provide clearly-labeled listing of:

1. Publications harvested that *have* already been cataloged by GPO, separating:
   a. Publications already cataloged by GPO, in both print and electronic format.
   b. Publications already cataloged by GPO based on print or microfiche editions in the FDLP, but now have an associated electronic file due to harvesting.
2. Publications, either electronic or print (or both), harvested that *have not* already been cataloged by GPO

Information for this report is derived from Contractor Task 8.

Deliverable Product # 8 Electronic dissemination to GPO of all information contained in all databases of all Harvested Content generated during this project. This may include either granting GPO complete access to, or the electronic delivery of, all information contained in these databases. This is an ongoing deliverable that should be continuously provided to GPO throughout the project.

## Deliverables, Time Line

1. Deliverable products #1-8 shall be submitted for review and discussion, prior to finalization and acceptance. The applicable stages are listed below.

   a. Step 1 - Contractor submits a draft of the deliverable product(s) to GPO.
   b. Step 2 - GPO reviews the draft(s).
   c. Step 3 - A follow-up conversation is held between GPO staff and contractor staff to discuss findings in draft report(s).
   d. Step 4 - Contractor makes necessary changes and issues Final Report(s).

2. The following charts provide suggested due dates for the various deliverable products. Contractors are encouraged to propose new time tables for each deliverable based on predicted timeframes. Please note that all deliverables must be met in a 180-day timeframe.

| | Deliverable Prod. # 1 | Deliverable Prod. # 2 | Deliverable Prod. #3 |
|---|---|---|---|
| Step 1 - Draft* | day 14 | day 43 | day 65 |
| Step 2 - GPO review** | within 3 days | within 3 days | within 3 days |
| Step 3 - Discussion*** | within 3 days | within 3 days | within 3 days |
| Step 4 - Finalization**** | within 2 days | within 2 days | within 2 days |

| | Deliverable Prod. # 4 | Deliverable Prod. # 5 | Deliverable Prod. #6 |
|---|---|---|---|
| Step 1 - Draft* | day 94 | day 116 | day 145 |
| Step 2 - GPO review** | within 3 days | within 3 days | within 3 days |
| Step 3 - Discussion*** | within 3 days | within 3 days | within 3 days |
| Step 4 - Finalization**** | within 2 days | within 2 days | within 2 days |

| | Deliverable Prod. #7 |
|---|---|
| Step 1 - Draft* | day 167 |
| Step 2 - GPO review** | within 3 days |
| Step 3 - Discussion*** | within 3 days |
| Step 4 - Finalization**** | within 2 days |

*Number of work days after the contract is awarded.
**Within the specified number of work days after Step 1
***Within the specified number of work days after Step 2
****Within the specified number of work days after Step 3

3. The delivery and acceptance completion date of 180 calendar days from the date of award. GPO will expect the project to be completed in 180 calendar days. The chart above maps out the due dates of deliverables on a 175 day period, with five extra days built into the schedule in order to allow for possible extenuating circumstances.