UNITED STATES GOVERNMENT PRINTING OFFICE (GPO)

WEB HARVESTING WHITE PAPER

V1.0

February 14, 2007

Table of Contents

1.0	**EXECUTIVE SUMMARY**	3
2.0	**INTRODUCTION AND BACKGROUND**	6
3.0	**VENDOR BACKGROUND AND TECHNOLOGIES**	8
4.0	**PILOT PROCESSES**	9
4.1	CRAWL AND RULE REFINEMENT PROCESSES	9
5.0	**PILOT RESULTS AND ANALYSIS**	10
5.1	CRAWL STATISTICS AND RESULTS	10
5.2	SCOPE DETERMINATION ACCURACY RATES	13
5.3	ANALYSIS OF DIFFERENCES AND DISCREPANCIES IN RESULTS	14
5.4	ACCURACY OF CONTENT HARVESTED	18
5.5	EXPLANATION OF COLLECTIONS COMPARISON AND ANALYSIS	19
5.6	HARVESTED METADATA ANALYSIS	21
5.7	GROUPING OF INDIVIDUAL DOCUMENTS INTO PUBLICATIONS	23
5.8	RULES CREATED FOR THE PILOT	24
5.9	ANALYSIS OF COMPARISON WITH MANUAL CRAWL	25
6.0	**LESSONS LEARNED**	26
6.1	TECHNOLOGICAL LESSONS LEARNED	27
6.2	RULE WRITING/PROCESS LESSONS LEARNED	27
6.3	POLICY LESSONS LEARNED	29
7.0	**RECOMMENDATIONS AND NEXT STEPS**	30
7.1	VENDOR RECOMMENDATIONS AND NEXT STEPS	30
7.2	GPO HARVESTING TEAM RECOMMENDATIONS	31
7.3	GPO HARVESTING NEXT STEPS	33

1.0 Executive Summary

In working toward the goal of building a comprehensive collection of content available through its dissemination programs, GPO plans to 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 harvesting technologies and methodologies, GPO has recently concluded a pilot project with two private companies (Information International Associates and Blue Angel Technologies) that harvested publications from the Environmental Protection Agency's Web sites, using rules and instructions that would determine whether 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. The overall results of the pilot are as follows:

Blue Angel Technologies Results

- Blue Angel identified 83,229 documents that were determined to be in scope and 1,845,592 documents that were determined to be out of scope at the end of the final crawl.

- Based on analysis conducted by GPO, Blue Angel's accuracy rate for determining content to be in scope was approximately 84%, and the accuracy rate for out of scope was approximately 70% at the end of the final crawl.

Information International Associates (IIA) Results

- Information International Associates (IIA) identified 239,478 documents that were determined to be in scope and 177,973 documents that were determined to be out of scope at the end of the final crawl.

- Based on analysis conducted by GPO, IIA's accuracy rate for determining content to be in scope was approximately 86%, and the accuracy rate for out of scope was approximately 64% at the end of the final crawl.

Summary of Findings and Lessons Learned

- The pilot demonstrated that scope determination of online documents discovered and harvested can be automated to a reasonable extent based on the accuracy rates found for each of the vendors' in scope populations, based on the general improvement of results from each crawl.

- The vendors reported a large variation in both the total number of documents discovered and the ratio of in scope documents to out of scope documents. These differences are due to the fact that each vendor discovered different populations of documents on the EPA Web site. Blue Angel discovered more content residing within databases, while IIA performed more targeted harvesting.

- Results from both vendors suggest that making rules too restrictive for in scope documents may result in the exclusion of large numbers of documents that may actually be in scope. The severity of this problem increases as the total population size of harvested content increases.

- Targeted harvesting will help GPO achieve higher accuracy rates. GPO should research each specific agency Web site to determine starting points and locations from which automated harvesting technologies will find the most in scope content. This can be accomplished through working directly with the agency's content experts and analyzing the agency Web domain. GPO will also develop a more targeted and detailed description of an online publication that falls within the scope of GPO dissemination programs (including the FDLP and C&I).

- Customization of harvester rules will be necessary for each site. According to vendor estimates at the end of the pilot, about half of the rules and instructions written in the pilot will need to be customized and tailored to each specific agency's content and terminology.

- In the current state, a great deal of additional manual processing will need to be performed by GPO after content is harvested in order for a given publication to be added to the FDLP collection. These processes include:

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

- Without a technological or financial solution to assist with the additional processing needed, an automated harvesting tool will only move the bottleneck from the discovery and harvest functions into the functions listed above. GPO will need to assess and analyze the business impacts of performing these processes with regard to time and resources available, and should explore ways of automating these processes.

- GPO will need to determine what functions and processes should be performed by the harvester. Certain tasks performed during these pilots (e.g., comparison of harvested content with GPO cataloging records, grouping of individual files into complete publications) can be performed using tools and technologies that may not be best performed by the harvester. GPO will need to assess these tools and determine the best workflow for these processes in the future.

Next Steps

GPO plans to implement a harvesting solution as a part of Release 2 of the Future Digital System (FDsys), currently scheduled to be implemented in mid 2008. In the meantime, the team recommends that GPO conduct another pilot (pending availability of funds) to further test methodologies and technologies. This pilot should include:

- Harvest of another test agency to be determined.
- Work with the test agency to identify publication characteristics and locations of possible in scope content.
- Identify and require specific starting points and parameters for the harvest, so that each vendor is targeting the same population of documents.
- Test of rules and instructions created for this pilot for their applicability to the new test site.
- A more detailed description of deliverables and objectives for the pilot, including minimum metadata, accuracy rates for scope determination and format of data deliverables.

GPO also plans to leverage its lessons learned with other similar projects being conducted by other agencies, including the NDIIPP initiatives and projects being conducted by DOE, NARA, Library of Congress, NTIS, and many others.

GPO has received all digital content from the pilot, and plans to conduct further analysis on the content, with the goal of cataloging in scope publications harvested from this pilot. Further analysis and preparation of the digital content will need to be completed before the content is ready for classification and cataloging. GPO intends to explore several cataloging methods, including automated metadata extraction, to create cataloging for material harvested from the pilot. GPO plans to provide permanent public access to the scope content harvested from the pilot that is cataloged. A plan to catalog harvested publications from this pilot that are within scope of GPO's information dissemination programs and not already represented in the Catalog of U.S. Government Publications (CGP) is being developed and will be reported at a later date.

2.0 Introduction and Background

2.1 Pilot History

A key mission of the U.S. Government Printing Office (GPO) is to provide permanent public access to all information products produced by the Federal Government. With the proliferation of online publishing, 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 (C&I). 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.

GPO has traditionally followed a manual process for discovery and harvesting of online electronic publications. Acquisitions staff and catalogers within GPO are involved in manual discovery of publications from various Federal Agency Web sites--using a Web browser to "point and click" to discover publications within scope. This is a very time-consuming and cumbersome process, and GPO has recognized the need to automate this process.

For the past few years, GPO has also used tools for harvesting content from the Internet to capture copies of *targeted* digital publications on Federal Agency Web sites. The publications reside dormant on an archive server until such time as the original version on the publishing agency Web site is no longer available. The harvested copy is downloaded and sent to an archive server and a PURL is assigned by GPO staff. GPO maintains full control of the harvested content and metadata in the archive and controls access privileges and mechanisms.

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

The proposed solution, referred to in this report as the harvester, will include discovery, assessment, and harvesting tools that will be used to harvest content. Discovery tools will locate electronic content from the 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 online versions of the content already exist in the system, and will establish appropriate relationships between versions. Harvesting tools gather the content and available metadata. The harvester technologies and the rules incorporated to discover and assess publications will be flexible so that they may be modified and updated as appropriate to harvest newly identified publications.

As a first step in learning about technologies and methodologies, GPO decided to conduct a pilot project with two private companies that harvested in scope publications from the Environmental Protection Agency's Web sites, using rules and instructions that would determine whether 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. GPO will

leverage the knowledge it acquires to build a set of requirements for the comprehensive harvesting solution to be implemented in conjunction with Release 2 of GPO's Future Digital System (FDSys).

2.2 White Paper Scope

It is important to note that this White Paper serves as the first of several updates from GPO with regard to Web discovery and harvesting. This paper reports on the project in 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 future direction and next steps for further development of the harvesting function that will be implemented during Release 2 of FDsys scheduled to be implemented in mid-2008.

As part of the update provided at the Federal Depository Library Council Meeting in April 2007, GPO will outline its plan to assess the content that has been retrieved from this pilot and catalog in scope publications to the FDLP collection where feasible.

2.3 Why EPA?

GPO chose the Environmental Protection Agency as the test Web site for this pilot for several reasons. Through manual discovery and harvesting processes, GPO has learned a great deal about the content published through the EPA Web site, and has identified many in scope fugitive publications. Also through this process, GPO recognized the opportunity to retrieve many publications not previously included in the FDLP.

Moreover, GPO has maintained a productive working relationship with EPA personnel, and EPA expressed great interest in participating in this type of pilot from its inception. Through extensive conversations with EPA, GPO was learned that as many as half of the publications on the EPA Web site have not been included in the FDLP. As a result, both GPO and EPA saw this pilot as a great opportunity both to identify fugitive publications on the EPA Web site and also to test harvesting tools and technologies on a group of Web sites that are known to contain a great deal of in scope content.

EPA's IT and information professional staff were actively involved in the project and have been very cooperative and supportive of this initiative. GPO is very appreciative of EPA's interest in the project and of their support for it.

2.4 SOW and Procurement Process

Since the main goal of the pilot project was to test and learn about different Web Harvesting methodologies and technologies available, GPO wanted to ensure that the pool of vendors targeted for the pilot was comprehensive and appropriate, and that any vendor who possessed these capabilities was able to bid on the project. GPO therefore issued a Request for Proposals (RFP) to the main contracting vehicle for the Federal Government, called FedBizOpps. The RFP included a statement of work (SOW) that explicitly stated the requirements of the project, and a complete set of criteria by which each of the vendors would be evaluated. For a complete copy of the SOW, please see Attachment #1.

The project generated a great deal of interest from the vendor community, and many good proposals were submitted. The proposals were evaluated by two separate boards of experts within GPO, a cost team and a technical team. Both teams reviewed all proposals and rated them based on evaluation factors set forth in the SOW and RFP.

In order to compare harvesting methodologies and technologies available, GPO decided to conduct the pilot using two vendors that would perform the tasks and deliverables simultaneously over the six-month period. To this end, GPO made awards to two vendors in early 2006.

3.0 Vendor Background and Technologies

3.1 Introduction to Pilot Contractors

GPO received a great deal of interest in the project, and many very good proposals were received. Of these proposals, two vendors were rated the highest by GPO's Selection Board in relation to the evaluation factors outlined in the SOW: Blue Angel Technologies (Blue Angel) and Information International Associates (IIA).

Blue Angel Technologies states that it "is an IT company based in King of Prussia, Pennsylvania, that specializes in the core business of providing government entities with standards-compliant, scaleable information-based solutions." The company demonstrated vast experience in providing Web harvesting services for other Federal Government entities, including the National Park Service, Defense Technical Information Center, National Oceanic and Atmospheric Association, as well as many state governments.

Information International Associates, Inc. (IIA) states that it "is an IT company based in Oak Ridge, Tennessee, that specializes in library and information management (LIM) and information technology services. IIA also employed a sub-contractor as a partner in the pilot, Digital Information Technologies, Inc. (DigInTech). DigInTech is a small specialty IT company that specializes in search and retrieval tools and techniques, crawler management, and hidden Web indexing. Both IIA and DigInTech have extensive relevant experience in working with similar projects, including projects with EPA, Department of Energy, and National Aeronautic and Space Administration."

3.2 Harvesting Technologies

Each vendor used different technologies to complete the discovery, assessment, and harvesting functions during the pilot.

3.2.1 Blue Angel Harvesting Technologies

Blue Angel's harvesting was accomplished by the MetaStar Harvester, Blue Angel's web crawler. MetaStar Harvester automates the steering, gathering, filtering, extraction/translation, mirroring, and consolidation of information from Web sites and other content sources. MetaStar Harvester was used for capturing and extracting data from the Web sites specified for this project, and was configured/customized to accommodate the rules for this project.

The Metastar Harvester crawls Web sites and directories of interest, extracts designated metadata and information, adds metadata (if appropriate), filters off undesired content, and consolidates the results in one or more organization-wide registers. For GPO's Web Harvesting Project, Harvester "hooks" were used to reflect the rules and instructions that the numerous out-of-the-box configuration settings do not address (e.g., custom steering and filtering algorithms). The set of hook-provided rules and instructions were encapsulated in documented code objects.

3.2.2 IIA Harvesting Technologies

IIA used a suite of tools available with Northern Light's Enterprise Search Engine (NLESE) Linux version 2.0. This product includes a configurable crawler that is capable of accessing any Web content and possesses robust reporting and monitoring capabilities. IIA also employed a suite of technologies provided by Autonomy to conduct the collections analysis task, which is described in Section 5.5 below.

4.0 Pilot Processes

4.1 Crawl and Rule Refinement Processes

Both vendors were given the same set of tasks and deliverables (as described in the SOW), and therefore followed the same overall process in completing each of the three crawls. For the first crawl, each vendor was provided with a document created by GPO that would serve as a starting point for rule creation and parameters for the project, entitled *Criteria and Parameters for GPO's Web Harvesting Pilot Project* (See Attachment #2).

With this document as a baseline, the same process was followed for each of the three crawls conducted during the pilot. First, each vendor created the rules to be used by the harvester to determine whether discovered content was in scope based on input from GPO. These rules were then delivered to GPO, who made suggestions, additions, edits, and comments on how the rules could be improved. Once the rules were approved by GPO, the vendor conducted the crawl of the EPA Web site and provided GPO with the results. GPO then conducted various analyses on the results of the crawl.

After analysis was complete, GPO then provided feedback, suggestions, and comments to each vendor on how well the harvesters were able to identify U.S. Government publications and make the determination of whether or not they were within scope of GPO's dissemination programs. A random sample was taken of documents that were determined to be in scope by the harvester and those that were determined to be out of scope. GPO cataloging and acquisition staff then looked at the random sample and determined if the scope determination made by the harvester was correct. This information was shared with the contractors, as well as the accuracy rate of scope determination for each. Based on the feedback given by GPO personnel, both vendors then refined their rules and instructions in preparation for the next crawl. This process was followed for each crawl during the pilot.

5.0 Pilot Results and Analysis

5.1 Crawl Statistics and Results

The following sections outline the overall statistics for each harvest conducted by each vendor. **It is important to note that all numbers of "documents" provided represent individual files and not necessarily publications. As will be discussed in later sections, there are many situations in which a publication is comprised of multiple files.**

5.1.1 Blue Angel Statistics and Results

	In Scope Documents	*%*	*Out of Scope Documents*	*%*	*Total Documents*	*Total Time to Crawl*
First crawl	12,115	6%	201,872	94%	213,987	24 days
Second crawl	43,395	7%	659,532	93%	702,927	36 days
Third crawl	83,229	5%	1,845,592	95%	1,928,823	43 days

As can be seen above, Blue Angel identified 83,229 documents it determined to be in scope at the end of the third crawl. The dramatic increase of documents found during each of the crawls reflects the addition of starting point URLs and the addition of content retrieved from databases (to be explained below). The following three charts provide more detail on the file types retrieved during each crawl.

Blue Angel First Crawl Results:

Document Type	In Scope	Out of Scope
HTML	6,438	173,892
PDF	5,250	23,338
Text	273	2,342
MS Word	115	1,151
MS Excel	38	848
MS PowerPoint	1	300
Total	**12,115**	**201,872**

Blue Angel Second Crawl Results

Document Type	In Scope	Out of Scope
HTML	17,295	50,1273
PDF	24,752	94,118
Text	1,095	59,902
MS PowerPoint	2	1,726
MS Excel	15	1,290
MS Word	236	1,223
Total	**43,395**	**659,532**

Blue Angel Third Crawl Results

Document Type	Qualifying	Excluded
HTML	59,859	1,705,303
PDF	22,389	104,958
Text	756	30,125
MS PowerPoint	0	2,126
MS Word	13	1,508
MS Excel	15	1,292
Total	**83,229**	**1,845,592**

As can be seen in the three charts above, the vast majority of file types that were retrieved by Blue Angel were either HTML or PDF files. Other formats found were MS Office files (Word, Excel, Powerpoint) and plain text files.

5.1.2 IIA Statistics and Results

	In Scope Documents	***%***	***Out of Scope Documents***	***%***	***Total Documents***	***Total Time to Crawl***
First crawl	121,911	35%	225,270	65%	343,848	N/A
Second crawl	108,067	35%	198,633	65%	312,134	N/A
Third crawl	239,478	57%	177,973	43%	423,449	N/A

The table above reports the number of documents collected by IIA's harvester for each crawl, and indicates whether the documents were in scope or out of scope. IIA also

identified documents that required “Special Handling,” or documents that need subjective human intervention to determine scope (the majority of these documents were documents that could be within the scope of C&I, but not the FDLP—see Attachment #2 for more details).

IIA identified 239,478 documents that it determined to be in scope, and 177,973 that were determined to be out of scope. The totals above do not include documents that were systematically deemed to be out of scope in the crawling and judgment process, based on conversation with GPO. The totals for the first crawl are higher than the second crawl because duplicate detection was not perfected until the second crawl. The time required for each crawl was not precisely estimated by IIA, mainly due to technological constraints experienced throughout each of the crawls. The duration of each IIA crawl, however, is estimated as about the same for each crawl as Blue Angel.

5.1.3 Database Crawling Methodologies and Results

The first crawl of the EPA Web site was scoped to include only “surface” Web pages in the EPA Web site, or pages on the site that can normally be reached by a Web crawler with no extra configuration. For the second and third crawls, GPO requested that each vendor write rules for and configure their harvesters to crawl several EPA databases selected by GPO. It must be noted that this was not a task that was specifically required in the SOW, but was requested as a first step for GPO in learning about methodologies available for database harvesting. It has not been GPO practice to harvest publications within databases heretofore. GPO has instead partnered with several agencies to provide access to database content.

The following databases were examined and crawled by the vendors:

- IRIS Risk Management Database (http://www.epa.gov/iris/)
- The Science Inventory ** (http://cfpub.epa.gov/si/)
- National Center for Environmental Assessment (http://cfpub.epa.gov/ncea/cfm/nceapubtopics.cfm?ActType=PublicationTopics)
- EPA National Publications Catalog
- RCRA Online (http://www.epa.gov/rcraonline/)
- CERCLIS (http://cfpub.epa.gov/supercpad/cursites/srchsites.cfm)
- ECHO (http://www.epa.gov/echo/)
- Envirofacts Data Warehouse (http://www.epa.gov/enviro/)

Both vendors examined each database and their specific characteristics before writing rules and configuring their harvesters to retrieve the content from databases. Several examples are below:

- Several query-based databases offered a “search all” or “retrieve all results” feature that can be activated by simply not entering any search terms or following conventions discovered by the vendors on how to retrieve all results. For these types of databases, the results pages were crawled and the same set of rules was applied to the content found.
- For databases that offered many fields for searching, the vendors configured their harvesters to follow conventions for retrieving all documents that met certain

requirements for fields. Once again, for these types of databases, the results pages were crawled and the same sets of rules were applied to the content found.

The results of the database crawling have not yet been specifically analyzed. Therefore, more analysis needs to be conducted regarding the success of the database crawling. Furthermore, GPO plans to make database harvesting a large part of the continued research and development required to build a comprehensive harvesting solution.

5.2 Scope Determination Accuracy Rates

The results of each crawl were provided to GPO by each vendor in database format. Scope determination accuracy rate was the main analysis conducted for each of the crawls. The objective of this analysis was to determine how accurate each vendor's harvesting technologies and rules were in determining whether discovered content was in scope for the FDLP and C&I.

GPO staff conducted the analysis between each crawl, reviewing a random sample of documents and determining whether the harvester scope determination was correct (an in scope and an out of scope population were provided to GPO for each crawl). GPO staff from the cataloging and acquisitions sections, who are experts in scope determination, conducted they analysis.

Since the total population of documents to review was so large, GPO extracted completely random samples of the documents retrieved. The sizes of these samples were relatively small for the first crawl (about 100 documents for each) because the level of accuracy was so low and the errors found were very consistent. GPO therefore decided that further analysis was not necessary for the first crawl. Sample sizes for both the second and third crawls were 500. While this sample size may seem relatively low compared to the overall population of documents retrieved, statistical analysis reveals that the percentages shown below are accurate within $\pm$4% for the second and third crawls. For example, for the 84% accuracy rate reported for Blue Angel in the third crawl, the actual accuracy rate for the entire population is between 80% and 88%.

Blue Angel Technologies

Scope determination accuracy rate	1st crawl	2nd crawl	3rd crawl
In scope	57%	68%	84%
Out of scope	92%	91%	70%

Information International Associates

Scope determination accuracy rate	1st crawl	2nd crawl	3rd crawl
In scope	42%	64%	86%
Out of scope	72%	38%	64%

As can be seen above, accuracy rates for content determined to be in scope by the harvester increased dramatically between each of the crawls. This improvement was based on feedback on scope determination errors provided by the same GPO experts that conducted the assessment and analysis. Based on this feedback, both vendors refined the rules, instructions, and harvester configurations for the next crawl.

The out of scope accuracy rates were in general less consistent than that of the in scope samples. This is mainly due to the introduction of more dynamic content (from databases, etc.) into the harvested content population in the second and third crawls. As stated above, much more work will need to be done to write rules specifically for these databases to make the harvester scope determination more accurate.

It is important to note that content scope determination is currently a manual process within GPO that relies heavily on human experts to make the scope determination. As a result, this process tends to be highly subjective. While the accuracy rate of the scope determination did not reach the 99% level initially set forth in the benchmarks as part of the SOW, GPO sees the improvement in automated scope determination demonstrated by both vendors during the pilot as a very positive start.

5.3 Analysis of Differences and Discrepancies in Results

As can be seen in the results reported in sections 5.1 and 5.2 above, the two vendors reported a wide variation in both the number of documents crawled overall and the ratio of in scope documents to out of scope documents. It is also highly evident that there is a large discrepancy in the results of both vendor pilots when the statistics presented by each vendor and the scope determination accuracy rates derived from GPO assessments are compared.

5.3.1 Difference in Total Number of Documents Found

There are several factors which explain the discrepancy in the results reported by each vendor. First, Blue Angel reported a much larger number of total documents found during the third crawl than Information International. One explanation of the large difference is that the numbers provided by IIA do not include documents that were systematically excluded in the scope assessment process because they were identified categorically as out of scope based on IIA's analysis and conversation with GPO. IIA conducted more analysis of content on the EPA Web site before each crawl, which enabled them to exclude categories of content and sections of the EPA Web sites that were known to be out of scope before harvesting them. In essence, this "targeted harvesting" methodology lessened the scope of content that needed to be assessed.

Conversely, Blue Angel implemented a different methodology for crawling the EPA Web site. Using URLs provided by GPO as starting points (identified by Blue Angel as "Seed URLs), Blue Angel crawled all URLs linked from these pages without excluding any specific areas within that could be excluded. Blue Angel was also able to crawl several selected databases more extensively during the third crawl than IIA, which yielded many more total documents crawled in the third crawl. This is not to say that IIA was unsuccessful in the delivery of the third crawl, but technical constraints experienced by IIA beyond their control prevented them from being able to extensively crawl several databases.

5.3.2 Difference in Percentage of In Scope Documents Found

Both vendors have reported a wide variation in the ratio of documents that are determined to be in scope to those deemed to be out of scope. While Blue Angel determined about 5% of their complete harvested population to be in scope for the third crawl, IIA determined about 57% to be in scope. Two factors help explain the difference in these results. First, based on the numbers reported by the vendors and the information presented above, the populations of harvested content retrieved by the vendors were drastically different from one another.

Second, the rules and instructions developed by each vendor were different. It is evident from the results that Blue Angel developed rules that excluded more total documents from the in scope population than IIA. However, it is also reasonable to expect that since IIA performed their crawls and applied their rules to a more targeted population of documents, a higher percentage of content retrieved would be in scope.

5.3.3 Analysis of Scope Accuracy Rates in Relation to Number of Documents Harvested

It is highly evident based upon the statistics presented by each vendor and the scope determination accuracy rates derived from GPO assessments that there is a large discrepancy in the results from the two vendors. In order to draw conclusions and make assessments of the results, GPO conducted further analysis of the results by comparing the statistics provided by each vendor with the scope determination accuracy rates determined by the GPO team. GPO used these figures to calculate the following estimated numbers for the population of documents harvested by each vendor:

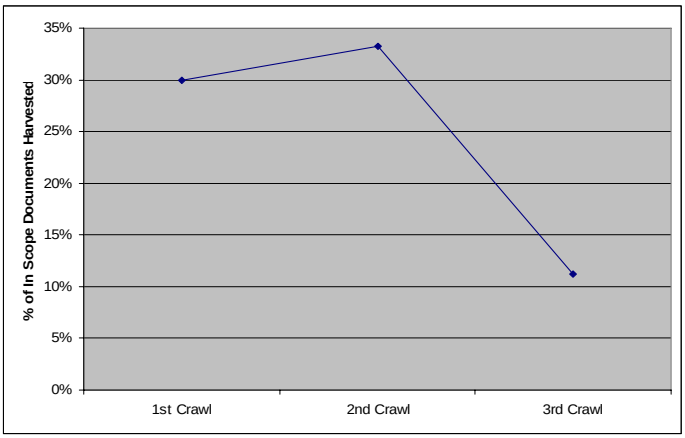
- total number of in scope documents for each vendor collection
- total number of documents that were determined by the vendors to be out of scope but were actually in scope
- ratio of number of in scope documents retrieved and correctly assessed for scope in relation to the projected total number of in scope documents in the harvested population

The following were the main findings of the analysis:

- Developing rules to be more exclusive in order to improve the in scope accuracy rate may eliminate a significant number of documents that are in scope.
- The impact on the number of in scope documents missed as a result of stricter rules dramatically increases as the total population size increases.
- Targeted harvesting will decrease the number of out of scope documents discovered.
- There may need to be a different rule creation and development process for dynamic content (e.g., database content) than for static content.

Blue Angel

The following chart outlines the percentage of total in scope documents correctly identified and harvested by Blue Angel. These percentages have been calculated based on the number of in scope and out of scope documents in relation to the corresponding accuracy rates determined by GPO staff.



As can be seen above, Blue Angel was able to correctly identify an estimated 30-35% of the projected total number of documents in the in scope population that Blue Angel harvested for the first and second crawls. This number dramatically decreased for the third crawl to about 10%. There are several reasons why this percentage changed so greatly between the second and third crawls. First, Blue Angel discovered and harvested many more documents (about 1.2 million) during the third crawl than the second, due to the introduction of new databases for the third crawl.

Coupled with this increase was an overall decrease in the accuracy percentage GPO found for Blue Angel's out of scope population from the second crawl to the third crawl (92% in the second crawl and 70% in the third). Therefore, since Blue Angel identified over 1.8 million documents that were deemed to be out of scope, an estimated 30% of these (540,000) are projected to be actually in scope.

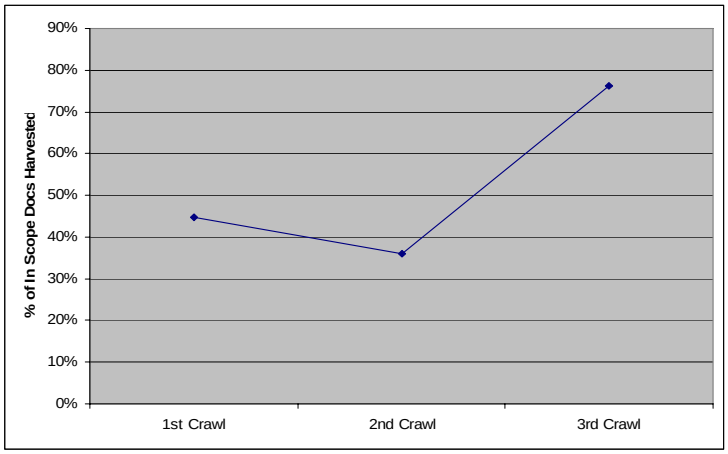
It must be stated that Blue Angel also added many rules and instructions that deemed documents that contained certain attributes to be out of scope, in order to achieve a higher rate of accuracy in identifying in scope material. Moreover, since many of the documents retrieved by Blue Angel during the third crawl were from databases, it is reasonable to state that there are many documents within the harvested population that have similar characteristics. Therefore, if a large number of documents from a database contained an attribute or keyword that was written as an exclusion rule by Blue Angel, there is a possibility that a large number of these documents were excluded as a result. This is evidence that:

1. Developing rules to be more exclusive in order to improve the in scope accuracy rate may eliminate a significant number of documents that are in scope.
2. The impact on the number of in scope documents missed as a result of stricter rules dramatically increases as the total population size increases.

3. There may need to be a different rule creation and development process for dynamic content (e.g., database content) than for static content.

IIA

The following chart outlines the percentage of total in scope documents correctly identified and harvested by IIA. Once again, these percentages have been calculated based on the number of in scope and out of scope documents in relation to the corresponding accuracy rates determined by GPO staff.



As can be seen above, IIA was able to correctly identify an estimated 35-40% of the projected number of documents in the in scope population that IIA harvested. This number dramatically increased for the third crawl to nearly 80%. There are several reasons why this percentage changed so dramatically between the second and third crawls. First, it has already been stated that IIA targeted specific areas of the EPA Web site to harvest, and excluded parts of the site hypothesized to be out of scope based on research and conversations with GPO.

Coupled with this increase was the addition of many rules and instructions in order to achieve a higher rate of accuracy in identifying in scope material. Moreover, there was a dramatic increase in the accuracy percentage GPO found for IIA's out of scope population from the second crawl to the third crawl (38% in the second crawl and 64% in the third). Therefore, since IIA identified only about 178,000 documents that were deemed to be out of scope in the third crawl, an estimated 35% of these (about 62,000 documents) are actually in scope.

This is further evidence that:

1. Developing rules to be more exclusive in order to improve the in scope accuracy rate may eliminate a significant number of documents that are in scope.

2. The impact on the number of in scope documents missed as a result of stricter rules dramatically increases as the total population size increases.

3. Targeted harvesting will decrease the number of out of scope documents discovered.

5.4 Accuracy of Content Harvested

Another form of analysis conducted by GPO after each crawl in the pilot was a test of how accurately the vendor tools were able to re-create the content harvested to have the same exact look, feel, and functionality of the content live on EPA Web sites. To accomplish this, GPO used the same sampling methodology as in the scope determination test. Starting with the samples, GPO staff manually compared the harvested content with the same content on the EPA Web site, looking for the following categories of error on the harvested content:

- *Missing text:* text was missing on the harvested file.
- *Incorrect text (including punctuation)*: text was incorrectly translated from the live copy to the harvested copy.
- *Missing images:* one or more images were missing on the harvested copy.
- *Incorrect images:* one or more images were not displayed correctly or in their entirety.
- *Content links within a publication were not local copies*: links that point to other URLs within the same publication were live EPA Web URLs instead of harvested copy URLs.
- *Faulty formatting of text and images:* Text or images on a harvested page were not formatted correctly.
- *Footer text missing:* the harvested copy was missing the footer text originally displayed on the Web copy.
- *Other (with explanation):* any other error, with an explanation as to what the error was.

The following is a comparison of accuracy for each of the vendors in each of the crawls:

Vendor	First Crawl Accuracy Rate	Second Crawl Accuracy Rate	Third Crawl Accuracy Rate
Blue Angel Technologies	48%	76%	80%
Information International Assoc.	30%	51%	60%

For the first crawl, the accuracy of both IIA and Blue Angel was very low. Most of this, however, can be attributed to technical difficulties experienced by both vendors during the first crawl.

Blue Angel's accuracy rate increased dramatically in the second and third crawls. In these crawls, over 80% of the errors found on the harvested content were errors that can easily be corrected, mainly missing text and missing images.

IIA's accuracy rate increased over the three crawls, but at a slower rate. This was most likely due to the technical difficulties experienced during the project. Most of the errors found were due to links being directed to live EPA pages rather than localized content. This issue is related to the grouping of partial publications (Section 4.6 below), and it is expected that many of these errors will be corrected when the methodologies for relating individual files within multi-part publications are further developed.

The team recognizes that more work will need to be done in order to test the feasibility of the harvested content to be cataloged and included in the FDLP. The plan for analysis of the content and its feasibility for inclusion in various dissemination programs will be discussed in the update provided for the next FDL Council Meeting.

5.5 Explanation of Collections Comparison and Analysis

In order to test the capability for automated tools to identify harvested content that is already cataloged by GPO, the final task to be completed by each vendor was collections comparison and analysis. In this task, both vendors were asked to perform a comparison where content retrieved from the Web crawler and other technologies were matched with MARC records created by GPO for the FDLP. The vendors were asked to:

a. 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.
b. 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.
c. Identify publications that have not been harvested or cataloged by GPO.

Both vendors have stated that positioning the task at the end of the pilot and not earlier in the line of tasks proved to be an obstacle in developing a custom solution and set of rules to complete the task as GPO had described. Consequently, both vendors have acknowledged that the results of the task are not as comprehensive as they had hoped.

GPO analysis of the success of this task is still in progress, and will be reported in the next iteration of this report in the specific context of GPO's plan moving forward to catalog and disseminate EPA content harvested during the pilot. This section focuses on methodologies and rules written to accomplish the task.

5.5.1 Blue Angel Collections Analysis Methodologies

Blue Angel performed this task examining each harvested publication in relation to each FDLP catalog record. A match was deemed to have been made if any one of the following five criteria were satisfied:

FDLP PURL Matches Harvested Publication URL: The technology resolved URLs given in the MARC catalog records supplied by GPO. If the resulting URL existed in the database of in scope documents then there was an exact match and the document was deemed to be a qualifying cataloged GPO publication.

Extended Title Match: Each catalog record title was calculated as the MARC subfield 245a element value, appended with a space, then appended with the MARC subfield 245b element value. The matching algorithm checked if a harvested qualifying publication had the exact same *Title* metadata field as a catalog record title. If such a match was found, then the document was deemed to be a qualifying GPO publication.

Extended Title Match with Stop Words: Each catalog record title was calculated as the MARC subfield 245a element value, appended with a space, then appended with the

MARC subfield 245b element value. The matching algorithm checked if a harvested qualifying publication had a *Title* metadata field *similar* to the title of a catalogued publication. Titles were deemed similar if and only if, after excluding all stop words (see Stop Word List below), the remaining words in each title were all present and in the same order. If such a match was found, then the document was deemed to be a qualifying GPO publication.

Proper Title Match: Each catalog record title was calculated as the MARC subfield 245a element value only. The matching algorithm checked if a harvested qualifying publication had the exact same *Title* metadata field as a catalog record title. If such a match was found, then the document was deemed to be a qualifying GPO publication.

Proper Title Match with Stop Words: Each catalog record title was calculated as the MARC subfield 245a element value only. The matching algorithm checked if a harvested qualifying publication had a *Title* metadata field *similar* to the title of a catalogued publication. Titles were deemed similar if and only if, after excluding all stop words (see Stop Word List below), the remaining words in each title were all present and in the same order. If such a match was found, then the document was deemed to be a qualifying GPO publication.

Blue Angel stated that the results were a bit disappointing, since the algorithm was only tested once on the last harvest and the benefit of adjusting the algorithm was missing. Although seemingly disappointing, without the benefit of a benchmark it is unclear to Blue Angel how many FDLP catalog entries are currently on the EPA site. They suggested that in the future it may make sense to include matching algorithms earlier in the project so that adjustments and confidence in the results can be established and the results subsequently improved.

5.5.2 IIA Collections Analysis Methodologies

The task of identifying duplicates and document versions presented several challenges for IIA. The first task was to identify documents that may have been cataloged (MARC collection) more than once primarily due to the document's existence in different formats or in different locations. IIA approached this task with the intent to compare only MARC record data and not to attempt to infer similarities from any electronic content that was available.

The second task was to identify harvested documents that have either versions or duplicates in different electronic formats or locations. For this, IIA harvested electronic document content from the GPO or EPA Web servers, then compared the results for identification of duplicates.

The methodology used for the tasks was to combine MARC record fields into three categories and to create "fusion" fields from them: Author-related, Title-related, and Subject-related. The premise for the approach to Author-related fields was an assumption that complex names of organizations may not have been fully controlled, complex roles of organizations (author, sponsor, owner, distributor, etc.) may have been vague at the time of cataloging, and personal authorship may have been cataloged with different spellings or formatting. Similarly, Title-related fields were assumed to include document title, series title, etc. The Subject-related fields are the closest to conceptual description of the documents and, in the absence of content, are the nearest possible

substitution for it.

IIA hypothesized that a conceptual or statistical match on any of these composite fields would have to happen when records described the same or closely related documents such as same document versions or same document in different formats. IIA did not initially fully appreciate that a match only on Author-related fields would tend to produce large numbers of false hits for different documents produced by the same organization or author. Furthermore, IIA relied on the ability to configure the software to attempt all the tests in sequence or combining the results: that is to identify documents that may be similar by more than one similarity test.

IIA fully recognizes that due to time constraints, they have not taken advantage of and tested finer similarities that could be attempted by comparing a few more specific fields and options available in the catalog, such as EPA publication number.

5.6 Harvested metadata analysis

GPO performed several tests to analyze the metadata retrieved for publications harvested during the third crawl. A completely random sample was taken of 500 documents from the third crawl to perform the following tests:

- ***Harvested metadata fields***: Analyzed which metadata fields were able to be retrieved for each metadata element for each document.

- ***Accuracy/correctness of "Title" metadata field.*** Compared the harvested "Title" metadata with the original document on the EPA Web site.

- ***Date Information.*** Looked at a subset of documents to see if date information is present on the pages and could be harvested in the next pilot.

While the SOW was not explicit on the desired metadata elements the harvester needed to retrieve, these tests have helped GPO gain an understanding of what descriptive and administrative metadata are available with documents on the Web. The following sections describe the results of each of these tests.

5.6.1 Harvested Metadata Fields

For this test the GPO team analyzed the metadata that were able to be retrieved for each document. The following fields were retrieved for 100% of the sample documents by BOTH vendors:

- File Format/MIME Type
- Retrieval or Harvested Date/Time
- File Size
- Time to Retrieve
- Checksum
- Resource URL
- Language

The following fields were retrieved for less than 100% of the sample documents by the vendors:

- Title (IIA: 99%, Blue Angel: 96%)
- Subject (IIA: 14%, Blue Angel: 8%)
- Author (IIA: 27%, Blue Angel 21%)

It is important to note that these were not the only metadata fields retrieved by the harvester, but were the ones that were deemed the most relevant to required descriptive or administrative metadata for content.

5.6.2 Accuracy/correctness of "Title" metadata field

Since "Title" was the only metadata field retrieved by the harvester that was deemed to be a useful descriptive field by GPO experts AND was retrieved for the vast majority of harvested documents, it was decided that this element would serve as the best test for determining accuracy of metadata for the documents. To perform this test, GPO compared the title reported as metadata by the harvester with the actual title or titles that appear on the publication.

The "title" metadata field was found to be about 75% accurate in capturing all available title information in the Blue Angel sample, while the same field was about 58% accurate in the IIA sample. It was noted by the evaluation team that many of the documents in the sample were individual documents within a series. Therefore, in most of these cases, the title of the series was retrieved, but not necessarily the title of the individual issue or document within the series. More than one title will need to be captured for these documents.

How GPO chooses to catalog publications within databases, following GPO Cataloging Guidelines, should drive how title information is captured. In some cases, GPO will want to identify all individual titles when each title is cataloged, but in other cases GPO will only need to harvest the collection level title. In some cases, the collection level title did not appear in the Subject field of the metadata as it was not specified in the rules that more than one title may be present and should be harvested.

It was also noted by the team that a significant number of documents within the sample were PDF files, which present an interesting challenge in retrieving title information. If a document was converted to PDF from another format (e.g., Quark, InDesign), it will most likely not maintain its original title metadata when converted.

5.6.3 Date Information

After an initial review of the metadata retrieved by both vendors, the GPO team was surprised by the lack of a "Date" field, besides the "Harvested Date". This was of specific concern to the GPO team because dates are often an indicator of versions and document modifications.

Based on this initial impression, GPO reviewed a small number of documents from the sample set to see if a date was present on the page that would be "harvestable." Based on the small number of documents studied, GPO found that about half displayed some sort of date that could offer an indication of version or date modified/updated.

GPO will need to study a larger sample of documents to see if modified/updated/created dates are present on a majority of the documents harvested in the pilot. In any case, future harvesting endeavors undertaken by GPO will need to be more specific on what required metadata elements the harvesting technologies should attempt to retrieve.

5.6.4 Conclusions from Metadata Analysis

The overall conclusion that can be drawn from the metadata analysis is that the amount of useful administrative and descriptive metadata retrieved from the pilot were minimal. Moreover, most metadata retrieved were more administrative in nature (e.g., file size, URL, etc.) than descriptive. In fact, the only descriptive metadata element that was deemed to be useful by the GPO team and that was consistently retrieved in the pilot was the title.

The low fidelity in available metadata was an expected outcome, given the logical way in which content is posted onto the Web. However, since GPO was not specific in describing the specific metadata elements expected to be retrieved by the harvesting technologies, it is not clear whether other useful metadata are still available on the Web that can be associated with publications. Therefore, it is important that GPO define exactly what metadata are required for the harvester to retrieve. Two sources that could be used as a basis for the required metadata are the minimum requirements for submission of content into FDsys and the minimum metadata requirements for MARC records.

5.7 Grouping of Individual Documents into Publications

It became clear during the pilot that there were many cases in which individual files harvested were not publications by themselves, but would only be determined to be in scope by GPO if they were grouped with the entire set of files which make up a particular publication. At the conclusion of the first crawl, GPO asked each vendor to write a set of rules that would begin to solve this problem. Once again, this was not an explicit task asked of the vendors in the SOW; therefore both vendors acknowledged that this task was a work in progress that would require a substantial amount of further development. While there are no hard numbers for the results of this task, a description of each of the methodologies developed by the vendors follows.

5.7.1 Blue Angel Publication Grouping Methodologies

Blue Angel approached this problem by defining a “compound publication”. A *compound publication group* was defined as a set of documents meeting all of the following conditions:

- The documents are qualifying (in scope) documents.
- Each document in the group contains a direct hyperlink to at least one other document in the group, or is directly hyperlinked from at least one other document in the group.
- The complete URL path (excluding filename and query string components) is the same for all documents in the group.
- Each compound publication group has a single *head document*, which is determined using the following steps:

- Eliminate from consideration all documents that contain any of the following words or phrases in the document: “app”, “appdx”, “appendices”, “appendix”, “appendixes”, “bibliography”, “ch”, “chap”, “chapter”, “citations”, “contributors”, “endnotes”, “exec sum”, “execsumm”, “glossary”, “illustration”, “legal notice”, “literature cited”, “notes”, “record of decision”, “ref”, “references”, “rod”, “section”, “summ”,
- If exactly one document remains, that document is the head publication.
- If other than one document remains (e.g. 0, 2, 3, etc.), the document that was first added to the Harvester queue is the head publication.

5.7.2 Information International Publication Grouping Methodologies

IIA acknowledged that its approach to packaging is still in the development stages. For most stand-alone HTML documents, all embedded elements (images, style sheets, importable JavaScript) were downloaded to a local copy and references modified. All in-server links found in what was considered to be the content body were localized, and linked documents were locally copied. Navigational links were converted to absolute addresses and associated with JavaScript producing a warning to the user leaving the GPO Web site. Local copies of documents, images, etc., were copied from the crawler archive when available and subsequently downloaded when needed. Some files were not available on the server due to the original files’ incorrect coding or because the files were permanently missing from EPA servers. Those cases could not be fixed by localization.

IIA were able to capture some valuable, derivable, and harvestable metadata that could be used in rules to identify hub pages and subordinate pages. Hub pages were localized together with immediately subordinate pages. In scope subordinate pages were provided with a cover page linking to the “parent” page.

5.8 Rules Created for the Pilot

The development of rules used by the harvester to determine whether content is within scope was a major component of this project. Both vendors spent a great deal of time developing the rules, based on their independent research, GPO’s Criteria and Parameters Document, and feedback provided by GPO throughout the project.

Of specific concern to GPO during the rule creation was the amount of customization that would be required to write rules for harvest of content from other agency Web sites in the future. Both vendors were asked to estimate what portion of the rules could be aggregated to harvesting other agencies. Both reported after the pilot that about half of the rules and instructions used in the pilot could be portable when harvesting other agency Web sites. The other half would require further customization; including name and text substitution (e.g., change “EPA” to “DOE”). The following two sections describe the methodologies used in rule creation by both vendors.

GPO’s review of the rules and results during the pilot indicate that the rules written and developed by Blue Angel are more exclusive than those of IIA. Blue Angel has fewer rules than IIA. Blue Angel had a more specific focus on specific types of file formats and known types of content while the IIA focus was more comprehensive through identification of publication categories from the universe of online official EPA

publications. There are advantages and disadvantages to both of these models, which will be discussed in the Lessons Learned section of this report.

5.8.1 Blue Angel Methodologies and Rule Creation

Two tests were applied by Blue Angel to determine if a document is in scope (e.g., considered to be an EPA publication). The first is to determine if the document is considered a publication, the second is to determine if the document is an EPA publication. Please see Attachment #3 for a complete list of the rules applied by Blue Angel.

5.8.2 IIA Methodologies Rule creation

To begin the rule-writing process, IIA developed categories for documents, observed samples of documents in these categories, and built the core set of rules from their observations. IIA intended to expand upon the classifications throughout the project, but found that accumulating documents by category was beyond the reasonable scope of the project. This core set of rules, after refinement in subsequent tasks, proved to be the basis for approximately two-thirds of the scope judgments. The core set of manually authored rules is the most portable, and should be useful as a starting point for crawls of other agencies.

IIA also developed new tools to help with the refinement process, including statistical analysis of rules performance and features of judged documents. IIA also included linked and linking documents in their rule development and scoring algorithm, and added function based rules. IIA also developed tools to automate their probabilistic rule creation. These tools revealed what IIA believes are the accuracy limitations of this approach to rule creation, which lies somewhere between 80–90%.

Please see Attachment #4 for a complete list of rules applied by IIA.

5.9 Analysis of comparison with manual crawl

It was important to determine what, if any, publications the automated harvesting tools did not harvest during the crawls. Subjective analysis was performed during the pilot by both GPO and the vendors to determine what publications were not harvested.

Using the third and final crawl harvest results, GPO compared the harvested publications from two EPA sub-agencies, the Office of Research and Development (<http://www.epa.gov/ORD/>) and the Office of Solid Waste and Emergency Management (<http://www.epa.gov/oswer/>), with over 100 MARC records representing online publications authored or co-authored by each of these two agencies.

These two sub-agencies were selected because GPO catalogers had manually reviewed and identified publications for harvesting from them. The URLs included in the bibliographic records were used as the point of comparison. GPO searched the harvested files for the file name of the URL. Almost all of the publications were monographs, and all were either in PDF or HTML file formats. The URLs were tested to ensure that they still were active and provided access to the specified EPA publications.

Of the publications from the Office of Research and Development, Blue Angel harvested 7.2% of the cataloged publications reviewed, and IIA harvested 15.8%. Of the publications from the Office of Solid Waste and Emergency Response, Blue Angel harvested 6.5% of the cataloged publications reviewed, and IIA harvested 25%. In a few instances, a HTML file format had been harvested whereas GPO had cataloged the PDF file format of the publications.

The results indicate that the rules of the harvesters certainly need additional review and modification. Although GPO retrieved a substantial number of files (between 1,834 and 4,669 files) from the harvests of both of these sub-agencies, between 75-93% of cataloged publications were not harvested. GPO plans to review the publications that were not harvested and where they are located on EPA Web pages to determine if there are similar characteristics GPO may identify and then incorporate into a future version of harvest rules.

This is another impetus for GPO to continue defining the characteristics of online publications, especially in this environment with increasing dynamic presentation of digital content, so that GPO may develop the most accurate in scope publication harvesting tools. GPO will also continue to work with agencies to help identify the nature and location of the online publications they publish. The harvester will be flexible to allow for modification of the rules as appropriate to help ensure the most comprehensive capture of publications possible. For any publications not harvested automatically, the more manual method of harvesting will continue.

GPO also plans to review the files from the vendors to determine if they represent in scope publications. Some of the files may represent parts of publications, such as chapters, and it is possible that some of the files do not contain content that GPO considers to be a publication. A cursory review of several of the thousands of files indicates that though the harvesters were able to identify and harvest numerous in scope publications, many more had been identified through manual review.

6.0 Lessons Learned

Overall, the GPO Team has concluded that the pilot has been valuable in learning about technologies and methodologies available for Web publication discovery, assessment, and harvesting. It is the unanimous view of the team that developing better harvesting tools will benefit GPO in the long term. While GPO will need to devote significant resources in developing them, the tools and processes implemented will ultimately lead to building a more comprehensive collection of content available through GPO dissemination programs.

The lessons learned and experience GPO has gained during the pilot will be instrumental in forming the plan forward developing these tools and technologies in conjunction with FDsys.

Both GPO and the vendors documented lessons learned and recommended next steps throughout the project. The lessons learned are summarized in the sections below, organized into the categories of Technological, Rule-writing/Process, and Policy Lessons Learned.

6.1 Technological Lessons Learned

- Both vendors stated that commercially available crawlers are not greatly differentiated and all are sufficiently scalable for GPO's needs. However, both vendors indicated that no commercial or free software meets all of GPO's specific requirements for document localization and packaging. It was therefore recommended by both vendors that GPO select a software for packaging that is highly customizable.

- Both vendors experienced technological constraints during the project that greatly increased the duration of the harvests. The vendors have provided GPO with full documentation on the technological constraints of each of the harvests. This documentation will need to be examined by GPO, and will need to be taken into consideration when GPO makes the policy decision of whether to purchase a harvesting solution and operate the harvesting function in-house or enlist the services of an external contractor for harvesting functions.

- The crawler configurations developed during this pilot were specific to the technologies used during the pilot. However, crawler instructions are sufficiently similar between products to be used by other harvesting technologies.

- Version control and duplicate detection will be key issues when GPO begins large-scale harvesting. Tools and technologies to accomplish this will prevent GPO from re-crawling content that was previously determined to be irrelevant or out of scope.

- As expected, publications in certain file formats, including PDF, MS Office applications, and text, were more easily harvested accurately than those in HTML or other file formats.

6.2 Rule Writing/Process Lessons Learned

- *Writing rules to increase the in scope accuracy rate may result in missing in scope content.*
    - GPO's ultimate goal for harvesting is to achieve both comprehensiveness and accuracy.
    - The results of the pilot show that both vendors worked to achieve a high level of accuracy for their in scope document populations (about 85%), but this compromised the accuracy rate for their out of scope document populations (65-70%)
    - As a result, an estimated 30-35% of the documents deemed to be out of scope by the vendors were actually in scope.
    - The impact on number of in scope documents missed as a result of this dramatically increases as the total population size increases.
    - Increased accuracy in automated scope determination could reduce the comprehensiveness of the identification and capture of online publications.

- *Customization of rules will be necessary for each target site.*
    - Some of the rules and instructions for this pilot were tailored specifically to identify and capture EPA publications. The vendors highlighted these

types of rules as they will not necessarily be applicable for harvest of all U.S. Government publications from all issuing agencies.
    - According to vendor estimates at the end of the pilot, about half of the rules and instructions written in the pilot will likely need to be customized and tailored to each specific agency's content and terminology.
    - These rules will also need to be updated regularly based on changes in environment of the site.

- *Targeted harvesting will help GPO achieve higher accuracy rates.*
    - GPO should research each specific agency Web site to determine starting points and locations from which automated harvesting technologies will find the most in scope content.
    - This can be accomplished through analysis of the Web site and working directly with the target agency's content experts.

- *Harvesting content from databases requires custom configurations and rules.*
    - Both vendors indicated that databases are rich sources of high-value/high-quality documents, and that database harvesting should be more prominent early in the project.
    - Each database requires a unique understanding of its organization and content. Gateway (or browse) pages can be an easy way to extract database objects fully and usefully.
    - Writing rules for and extracting in scope documents from databases is unique to each particular database. Little can be generalized from one database to the next.

- *Collections Analysis*:
    - The vendors reported that since the collections analysis algorithm was only run once during the pilot, the benefit of adjusting the rules was missing.
    - Without the benefit of a benchmark, it is unclear as to the number of publications on the EPA Web site that have been cataloged for inclusion in the FDLP.
    - GPO will need to conduct more analysis on the results of the collections analysis portion.
    - In the future, GPO should include algorithms such as the matching earlier in the project so that adjustments and confidence in the results can be established and the results subsequently improved.

- *As expected, it has been difficult to completely automate the scope determination process, which has traditionally been largely subjective.*
    - Some publications identified within a rule locating publications by category type (e.g. conference or meeting proceedings or notes or draft publications which are "final" publications such as Draft Environmental Impact Statements or internal works in progress) are within scope publications and some are not. The challenge remains to identify the characteristics of those within scope and those that are not.

- *Publication metadata provided some but not all cues for determination whether the file being analyzed is an in scope publication.*

- Terminology, links, and other cues within the file also assisted in the identification of publications.

6.3 Policy Lessons Learned

- As the accuracy of harvesting tools increases, the comprehensiveness of the harvest could be compromised.

- A great deal of additional processing will be required after content is harvested in order for the publication to be added to the FDLP collection, including:
  - Grouping of individual documents into logical publications: both vendors attempted to write rules to automate this process, but more work will need to be done to improve the precision of these tools.
  - Inspection of harvested content: manual human review of content harvested for scope determination.
  - Cataloging: creation of cataloging records for in scope content.

- It is essential to work with an agency's Web support personnel to obtain the information necessary to normalize different representations of the same URL. Harvesting an agency website requires extensive interaction with agency personnel in order to disambiguate alternate representations of URLs, and to identify starting points sufficient to ensure adequate coverage.

- Given the amount of content that will need to be reviewed for scope as a result of automated harvesting, GPO will need to make the decision whether to include un-reviewed (and possibly out of scope material) in the CGP.

- GPO should anticipate allocating more time identifying publications outside of agency Web site domains for some agencies that widely use commercial web hosting and/or hosting at research institutions for their specialized content and applications (e.g., DoD, DOE).

- Documents that are clearly out of scope tend to affect the automated judgment process. GPO should consider identifying areas of Web sites that include a large number of in scope documents early in the process, and focusing on these areas.

- Both vendors indicated that GPO should allocate more project effort to harvesting database documents, and schedule this effort earlier in the project. This process will inevitably require extensive interaction with agency personnel and substantial lead time.

- Several publications at EPA are in databases that use robot exclusions. Per GPO policy, we will not harvest these publications using automated harvesting. (These publications may be available elsewhere on the EPA Web site in other publication repositories.)

- Increasing recall of the discovery tools complicates development of publication assessment tools. Developing more precise tools to identify, assess, and harvest content more easily as in scope will sacrifice recall for now; however, the tools will be more precise and harvest more content that includes U.S. Government publications.

This will save GPO time and effort managing, classifying, and cataloging appropriate content.

7.0 Recommendations and Next Steps

Based on lessons learned throughout the pilots, both GPO and the vendor teams have compiled a list of recommendations and next steps for the development of discovery and harvesting tools in conjunction with FDsys.

7.1 Vendor Recommendations and Next Steps

The following is a list of recommendations received from both vendors:

- Establish Regression Database: It is recommended that a database of known in and out of scope documents be established so that changes to rules can be tested for impact before deployment.

- Increase Confidence: Rules should be statistically tested before being deployed and confidence in their effectiveness gained.

- Enhance Document Structure Analysis: Based on the increasing sophistication of rules examining the type and structure of documents, it is recommended that the Harvester be configured with an enhanced document analysis capability. By adding a module that literally dissects each document, a number of enhanced rules can be introduced that more accurately determine the type of each document and thus whether or not it is in or out of scope.

- Enhance Publication Grouping: It is recommended that a more in depth study be conducted into compound publications. By examining a greater number of compound publications it will be possible to develop more sophisticated rules which result in increased accuracy.

- Enhance Draft Detection: It is recommended that the Harvester be enhanced to better detect draft documents. Currently drafts are detected textually by the NP_Draft rule which checks for the word draft in several metadata fields. It is recommended that the harvest be enhanced to detect graphical draft watermarks.

- Localize Content Messaging: Localized content would be better suited for publication if each localized content page contained a message indicating that it is a harvested page, and when it was harvested. This feature was deemed to be outside the project scope.

- Collections Analysis: A cursory analysis of the matching results showed that a number of duplicate matches occurred as a result of a matching title which in the case of HTML is very often incorrectly populated. Another issue that was observed was that a number of URLs in the GPO provided records resolved to a message indicating that the publication had been archived and was not available for comparison. In the future it is recommended that the Matching algorithm be modified

to compare a more extensive amount of document content instead of only the metadata as is currently the case.

7.2 GPO Harvesting Team Recommendations

Based on lessons learned from the pilot and recommendations provided by the vendors, the GPO harvesting team recommends the following:

- GPO will need to address the issue of what is the most effective methodology for harvesting--accuracy or comprehensiveness (precision or recall). Results from the pilot indicate that increased accuracy could reduce the comprehensiveness of the identification and capture of online publications, but if harvester rules are relaxed to include more content, it is possible that a large number of out of scope documents will be harvested.

- The overall long term goal for harvesting is to achieve both maximum accuracy and comprehensiveness, but given current resource constraints in being able to process harvested content, the GPO harvesting team agrees that the short term goal should be accuracy. The ultimate goal is an accurate harvest that is fully comprehensive, identifying all online publications in scope.

- The pilot demonstrated that scope determination of online documents discovered and harvested can be automated to a reasonable extent based on the accuracy rates found for each of the vendors' in scope populations and the general improvement of results from each crawl. However, GPO will need to decide what the acceptable rate of accuracy is for automated scope determination. The GPO team was pleased overall with the amount of improvement demonstrated by each of the vendors in automated determination, but believes that more work should be put into this effort to increase accuracy and identify valuable in scope publications.

- However, discovery and harvesting are just two parts of the overall acquisition and cataloging function. Several processes will need to take place before the content can be added to the FDLP collection, all of which are currently manual processes:
    - Grouping of portions of documents into entire publications: both vendors attempted to write rules to automate this process, but more work will need to be done to improve the precision of these tools.
    - Inspection of harvested content: manual human review of content harvested for scope determination.
    - Cataloging: creation of cataloging records and classification for in scope content.

- While the ability to automatically discover and harvest many documents in one crawl is a significant achievement, the current staffing and funding within GPO falls well short of being able to process this volume. Without a technological or financial solution to assist these other legacy processes, all that an automated web harvesting tool will do is move the bottleneck from the discovery and harvest functions into the classification and cataloging functions. GPO will need to assess the impact of performing these processes with regard to time and resources available, and make a policy decision on what functions should be performed by the harvester and what can be automated in the future.

- GPO will need to make the decision whether to acquire and develop harvesting tools in-house or enlist the services of a contractor to perform harvesting functions. Based on technological constraints expressed by both vendors, the team recommends that harvesting services be contracted. If GPO had the right expertise and infrastructure, it could be done in-house, but it will most likely be more cost effective to have a contracted solution.

- In order for automated Web discovery and harvesting to be successful, it is the team's position that GPO engage in extensive collaboration with the publishing agency. GPO will need to work with both Web masters and information professionals within the agencies to obtain a better understanding of their content, the characteristics of their specific in scope publications, and the location and nature of the content they produce. This will greatly assist in the rule-writing process and will greatly reduce the amount of out of scope content harvested.

- Some of the rules and instructions of the initial harvesting solution may limit some category types of publications identified and harvested to reduce the number of out of scope publications retrieved. Knowing these limits will provide a starting point for manual review of Web sites.

- GPO will need to identify starting points, including databases and other dynamic Web pages, for each agency Web site that will be harvested in the future. These starting points should include known publication repositories.

- This pilot included a one-time comparison between the results of the third crawl and records in the CGP. Further testing of this process, which will enable GPO to identify what still needs bibliographic control, should be conducted, as both vendors indicated that they noticed some problems in the comparison.

- GPO will need to leverage its lessons learned with other similar projects being conducted by other agencies, including the NDIIPP initiatives and projects being conducted by DOE, NARA, Library of Congress, and many others.

- GPO should consider a more detailed structural analysis of documents to help identify and determine the relationship between partial and compound publications.

7.3 GPO Harvesting Next Steps

GPO plans to implement a harvesting solution as a part of Release 2 of the Future Digital System (FDsys), currently scheduled to be implemented in mid to late 2008. In the meantime, the team recommends that GPO conduct another pilot (pending availability of funds) to further test methodologies and technologies. This pilot will include:

- Harvest of another test agency to be determined.
- Extensive work with the test agency to identify publication characteristics and locations of possible in scope content.
- Test of rules and instructions created for this pilot for their applicability to the new test site.
- A more detailed description of deliverables and objectives for the pilot, including minimum metadata, accuracy rates for scope determination and format of data deliverables.

While the new pilot is being conducted, GPO also plans to leverage its lessons learned with other similar projects being conducted by other agencies, including the NDIIPP initiatives and projects being conducted by DOE, NARA, Library of Congress, NTIS, and many others.

GPO has received all digital content from the pilot, and plans to conduct further analysis on the content, with the goal of cataloging in scope publications harvested from this pilot. Further analysis and preparation of the digital content will need to be completed before the content is ready for classification and cataloging. GPO intends to explore several cataloging methods, including automated metadata extraction, to create cataloging for material harvested from the pilot. As appropriate, GPO will provide permanent public access to the publications harvested from the pilot. A plan to catalog harvested publications from this pilot that are within scope of GPO's information dissemination programs and not already represented in the CGP is being developed and will be reported at a later date.

GPO will also develop a more targeted and detailed definition of an online publication that falls within the scope of GPO dissemination programs (including the FDLP and C&I). As part of this effort, GPO will look at characteristics of publications that were not harvested during this pilot to assist in rule writing and definitions of in scope content.

A list of the types of documents a harvesting tool may identify that are not publications within scope of GPO's information dissemination programs is also being considered to support future development of publication harvesting technologies. Moreover, while automated harvesting technology solutions improve to allow for both accuracy and comprehensiveness, GPO will continue to identify and manually harvest publications not captured in an initial automated harvest solution.

Attachment #1: Web Harvesting SOW

Attachment #2: *Criteria and Parameters for GPO's Web Harvesting Pilot Project*

Attachment #3: *Blue Angel Rules*

Attachment #3: *IIA Rules*

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.

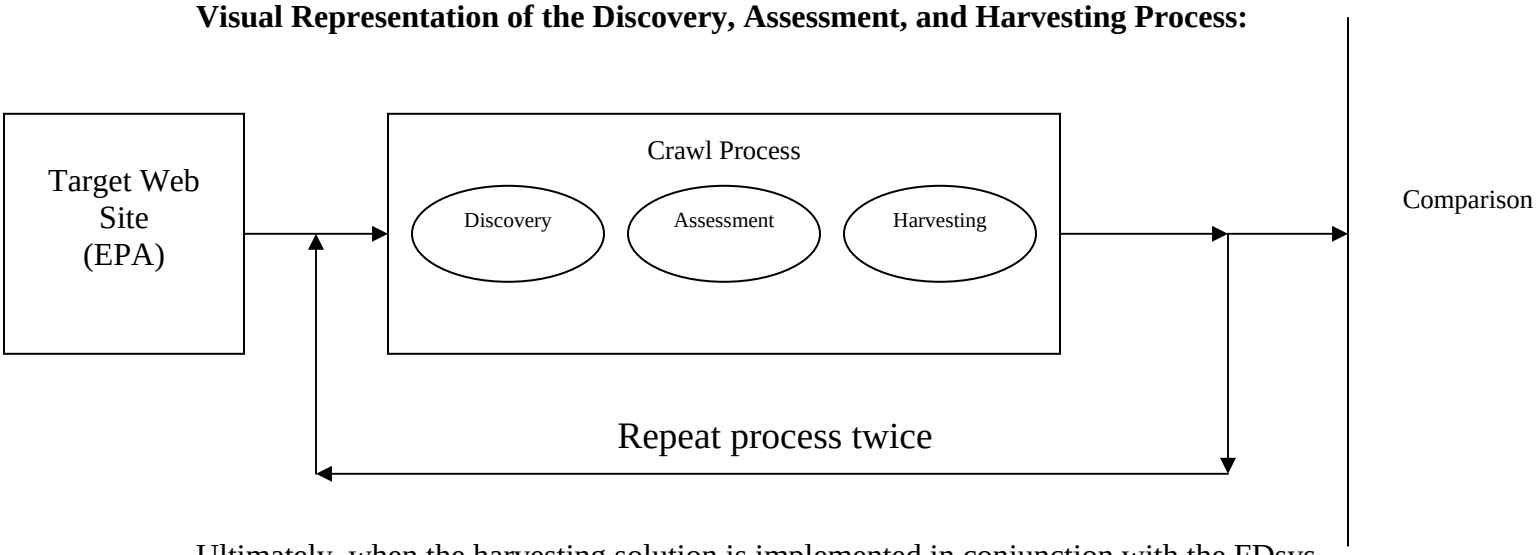
- *Metric:* Accuracy of publications located by the crawler
    - By the end of the third crawl, a maximum of 10% error between the **number** of in-scope documents harvested by the crawler technology and that of benchmark #1 (manual harvest conducted by GPO staff).

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.

CRITERIA AND PARAMETERS FOR

GPO'S WEB HARVESTING PILOT PROJECT

TABLE OF CONTENTS

1 INTRODUCTION 3

2 SCOPE OF PUBLICATIONS TO BE HARVESTED 3

3 DEFINITION OF GOVERNMENT PUBLICATION 4

4 ATTRIBUTES OF ONLINE PUBLICATIONS TO BE HARVESTED 6

5 ATTACHMENTS 11

5.1 Examples of publications 11

5.2 Examples of published information not considered publications 11

5.3 Publication terminology in English 12

5.4 Publication terminology in Spanish 15

5.5 File extensions in GPO's Catalog of U.S. Government Publications PURL server 17

5.6 Other file extensions 17

1 INTRODUCTION

This document outlines criteria specifying the characteristics of publications within scope of GPO's information dissemination programs and the pilot project to harvest publications from the U.S. Environmental Protection Agency (EPA) Web site. The crawler technology rules, instructions, and parameters should capture all EPA publications meeting these criteria so that the U.S. Government Printing Office (GPO) may provide permanent public access to them through its information dissemination programs.

GPO is looking for publications and any associated metadata within scope of the Federal Depository Library Program (FDLP) and the National Bibliography of U.S. Government Publications. Definitions of these two programs follow in this document.

EPA publications and their associated metadata to be harvested in this pilot project are those that EPA publishes, disseminates, or makes available to the public. These publications can be in any language, in any form or format, and in any location on official Web pages, including deep Web sites.

As outlined in Contractor Tasks #1 and 2 in the Statement of Work (p. 7), GPO will collaborate with the contractor to develop rules, instructions, and crawl parameters. The attributes of online publications listed in section 4 of this document are not prescriptive but are meant to serve as a basis for discussion about the rules, instructions, and parameters to be employed.

2 SCOPE OF PUBLICATIONS TO BE HARVESTED

Publications to be harvested are those issued by EPA and within scope of the Federal Depository Library Program and the National Bibliography of U.S. Government Publications.

Scope of the Federal Depository Library Program

The scope of the FDLP includes all published Federal government information products, regardless of format or medium, which are of public interest or educational value, except for those products which are for strictly administrative or operational purposes, classified for reasons of national security, or the use of which is constrained by privacy considerations.

Included in the FDLP are publications created as a result of U.S. Government funded contract or grant. Included in the front matter of these publications is a statement indicating that the publication was funded by a grant or contract or produced under contract or grant. Publications funded through grant or contract may have more than one issuing agency, the Federal agency and another publisher. Publications at the National Sea Grant Library at

http://nsgd.gso.uri.edu/ are U.S. Government publications as the funding for the publications is provided by the National Oceanic and Atmospheric Administration.

Scope of National Bibliography of U.S. Government Publications

The National Bibliography includes all publications in the FDLP as well as cooperative publications and other U.S. Government publications that are for strictly administrative and/or operational purposes (e.g. forms).

The National Bibliography is a comprehensive catalog containing descriptions and locations of U.S. Government unclassified publications in all formats. The National Bibliography describes any publication, regardless of form or format that any U.S. Government agency publishes, disseminates, or makes available to the public that is of public interest or educational value, as well as any publication produced for administrative or operational purposes. Publications represented in the National Bibliography are acquired from official sources or sites, and are subject to official use or security classification restrictions.

In short, the National Bibliography is a "comprehensive index of public documents," including "every document issued or published" not subject to official use restrictions or "not confidential in character". Source: 44 U.S. Code §1710 http://www.access.gpo.gov/uscode/title44/chapter17_.html

Publications identified for the National Bibliography are cataloged and appear in the GPO's Catalog of U.S. Government Publications. A new version of the catalog is currently in development at http://franklin.gpo.gov/.

It is presumed that information accessible on an agency's public Web site is not for strictly administrative or operational purposes, classified for reasons of national security, or constrained by privacy considerations. It is also presumed that some cooperative publications may be publicly accessible online (and the issuing Federal agency recovers costs by selling the tangible format). Therefore, all publicly accessible publications on the Internet that EPA has published, disseminated, or made available to the public should be harvested.

3 DEFINITION OF GOVERNMENT PUBLICATION

"**Government publication**" means informational matter which is published as an individual document at Government expense, or as required by law. Source: 44 U.S.C. §1901 http://www.access.gpo.gov/uscode/title44/chapter19_.html

Additional clarification of the definitions

A government publication is a work of the United States Government, regardless of form or format, which is created or compiled in whole or in part at Government expense, or as required by law.

In this pilot, an EPA publication to be harvested must be a publication that EPA publishes, disseminates, or makes available to the public and is from official sources or sites. Online U.S. Government Web sites typically have .gov, .mil, or fed.us domains; however, other domains, including .org, .edu, and .com, are also used at some official Web sites. Any publications on Web sites operated by an entity other than EPA but under Federal contract or grant by the EPA should be harvested as the publications therein are official EPA publications. EPA publications reposted on unofficial Web sites where EPA is not responsible for the posting as the official issuing agency should be excluded.

Different versions or editions of monograph or serial publications are separate government publications.

Publications include, but are not limited to, books, newsletters, journals, pamphlets, maps, and video recordings. They also include other published information such as **some** news releases and application forms. They may also be entire databases, PDF files, or MS Excel spreadsheets.
Examples include:

- U.S. Copyright Office Factsheets http://www.copyright.gov/circs/index.html#fl (Pamphlet-like publications)
- Agricultural Outlook: statistical indicators http://purl.access.gpo.gov/GPO/LPS50465 (Largely comprised on Excel spreadsheets)
- ERIC http://purl.access.gpo.gov/GPO/LPS54302 (Replaced previously published print publications that were indexes to U.S. Department of Education journal literature.)
- Producer Price Indexes http://purl.access.gpo.gov/GPO/LPS58465 (Publication with "news release" in the title but a publication longer than a one-page media release.)

Publications may also be integrating resource. An integrating resource is a "bibliographic resource that is added to or changed by means of updates that do not remain discrete and are integrated into the whole. Integrating resources can be finite or continuing. Examples of integrating resources include updating publications updated by loose-leafs and updating Web sites." (*Anglo-American Cataloging Rules*, 2002 Revision) They are publications that do not retain discrete parts. When they have an update, the update is incorporated into the whole. These may be basic manuals that are updated by separately published changes, transmittals, amendments, etc. Some publications of this type have separate updates that are not interfiled into the basic volume but are separate from the main publication. Others, called looseleaf when in print, have update pages that are interfiled into the main publication. GPO receives and catalogs many of these kinds of publications, and each is one publication. Examples include:

- International Flight Information Manual http://www.faa.gov/ats/aat/ifim/index.htm (Integrating resource)
- H.I.P. Pocket Change http://www.usmint.gov/kids/flashIndex.cfm (Integrating resource)

Deep Web databases that include separate monograph or serial publications should be crawled and each separate publication therein should be harvested.

Title 44 *U.S. Code* uses the word “document”. For the purpose of the pilot project, the use of “document” and “publication” above are synonymous. GPO prefers the term “publication”.

Fugitive publications

It is assumed that many of the publications to be harvested are currently “fugitive publications”. A fugitive publication is a U.S. Government publication that falls within the scope of the FDLP and/or the National Bibliography, but has not yet been identified/obtained and included in the information dissemination program(s). Once identified, fugitive publications are added to the National Bibliography and, if in scope, made accessible to the FDLP. Fugitive publications usually occur when Federal agencies publish on their own, without going through GPO. These publications may include tangible products, but they most commonly now are publications posted online only. Fugitive publications may be located in deep Web sites, where identification of publications has proven to be complicated.

4 ATTRIBUTES OF ONLINE PUBLICATIONS TO BE HARVESTED

The following list is organized by category for reference purposes and convenience. The categorization does not imply any ranking.

Location

EPA publications are located in EPA official sources or sites.

Publications are most likely within the http://www.epa.gov domain and sub-domains.

Publications may be outside of http://www.epa.gov. Web pages with different domains than www.epa.gov (primarily found through the EPA Web site index) include, but may not be limited to:

- http://www.bwc.gov/ (joint project with the U.S. Department of Transportation)
- http://www.energystar.gov/ (redirects from www.epa.gov/energystar/)
- http://www.ert.org
- http://es.epa.gov/
- http://nepis.epa.gov/
- http://cfpub.epa.gov/ncea/
- http://es.epa.gov/ncer/
- http://yosemite.epa.gov/

Publications are located throughout EPA Web sites, including but not limited to:

- Deep Web sites
- Query-based databases

- Agency content management systems
- Dynamically generated Web pages
- On FTP servers
- Behind proxy servers
- Behind firewalls

Publications may be located through page links. GPO recommends the following:

- Crawl all pages of the EPA Web site in order to locate and harvest all in-scope publications.
- Weigh .gov, .mil, .fed, and .us higher when linking to pages outside of the EPA.gov domain.
- Stop a crawl thread when a boundary indicator" (such as exit signs or scripts) is present, but ONLY when the page being linked to does not contain official Federal information.

Metadata

Publications will have metadata associated with them, which must be captured along with its entire corresponding publication. Metadata includes such information as:

- Title and caption
- Author, Creator, Publisher, Authority or Rights Owner (i.e. the agency's name or abbreviation)
- Provenance
- Resource type or Description (indicating the resource is a "publication", "document", "text", or related term)
- Version, fixity, and relationship to other publications
- Technical, structural, file format, packaging and representation information
- Administrative information

Parameters

Publications may have other information, objects, or applications associated with them that are required to render the harvested content accurately. The harvester must capture and harvest all such information.

The crawler should harvest entire publications. In some instances, there may be publications that are posted as HTML Web pages with hyperlinks rather than PDF files. The crawler must harvest all Web pages that comprise the publication and ensure that all hyperlinks are correct and valid.

Publications not issued by EPA are not within the scope of this pilot project. For example, the EPA posts sections from publications, such as the *Federal Register* and the *Code of Federal Regulations,* issued by other Federal agencies on its Web site. These are not authored and issued by EPA.

Publication identification

Known major EPA publication sources include:

- EPA Publications Source
  http://www.epa.gov/epahome/publications.htm
- National Environmental Publications Information System
  http://nepis.epa.gov/
- Foreign language publications
  http://yosemite.epa.gov/ncepihom/nsCatalog.nsf/foreign?openform&CartID=12776-020558
- Newsletters list. EPA Newsletters at
  http://www.epa.gov/epahome/newslett.htm have irregular publication cycles. EPA does not publish journals

Proper nouns, including an agency name, publication title, author name, and author affiliation, in the first 250 words on a Web page indicate the beginning of a text block, which is likely to be part of a publication.

The Federal agency name located in the front matter or last ten pages of a several page document help to identify a publication, especially when on an agency server and/or when “authored by” or “authors” is located near the agency name. These are more likely to be publications in scope (published by the agency) than publications by another author about the agency. The beginning and ending pages in a publication typically include bibliographic and agency author (statement of responsibility) information.

An ISBN or ISSN, especially in the front matter or last ten pages of a publication or in the metadata, often identify a publication from other types of information on Web pages.

Information about publications and the publications themselves include common words or phrases that describe publications. See Attachments 5.3 and 5.4 for publication types and trigger words that typically are found in or near links to publications. The greater the number of these words together, the greater the likelihood the file is a publication.

Web pages including publications may have information in running headers and footers that specify the publication or chapter titles, statement of responsibility (agency author information), or other publication information, such as report numbers.

Publications may be available in different versions, which should be identified through the metadata. If change information is not in the metadata, other possible version triggers include but may not be limited to:

- Modifications to the content
- Changes to the “last updated” date
- Language translations
- Changes to a publication’s title
- Changes to a publication’s edition statement
- Changes in the issuing agency of a publication

- Changes in file format (e.g., TIFF to JPEG)
- Levels of authentication (e.g., authentic vs. official)
- Changes to the publication's numbering (e.g. volume 100, issue 50, year 2005, etc.)

The following, along with text, are considered part of a publication:

- Embedded files
- Background graphics
- Java applets
- Audio and video

File formats

Publications will be available in all types of file formats, including but not limited to:

- PDF
- HTML
- Audio
- Video
- Dynamic content
- Proprietary word processing software
- Rich media
- XML

Per EPA, all but a few older PDFs are 508c compliant. Newer PDFs may be broken up into several smaller files. See Attachments 5.5 and 5.6 for the most common file types (from all Federal agencies) found in the Catalog of U.S. Government Publications in March 2005.

The same publication may be available in more than one file format. For example, a publication may be disseminated in PDF, Word, and HTML. In some cases, the publications are identical in each format, but in others, one format may, for example, contain additional functionality and/or content. All file formats should be harvested so that the assessment tool and GPO catalogers may evaluate any differences between the formats.

Other

Publications that include statements in the front matter indicating that the document or publication was funded by grant or contract are official U.S. Government publications.

Publications that are only partially harvested by the automated harvester should be flagged and time stamped for manual follow-up and special review by GPO Staff.

A publication that is inaccessible because it is available through a login and password may be a cooperative publication. Place information about these publications in a separate folder from other results for special review by GPO staff.

Publications including a copyright statement < © copyright > in the front matter stating that copyrighted material is included in the publication may be a cooperative publication. Place these publications in a separate folder from other results for special review by GPO Staff.

Publications including the following words or phrases in the front matter or end of the publications or in the metadata may be within scope of the National Bibliography but not within scope of the FDLP. We ask that you identify the following groups by placing them in a separate folder from other results for special review by GPO Staff.

- For official use only
- For internal use only
- For administrative use only
- For operational use only

Publications including the following words or phrases in the front matter or end of the publications or in the metadata may have been inadvertently posted on the public Internet. We ask that you identify the following groups by placing them in a separate folder from other results for special review by GPO Staff.

- Restricted
- Classified

5 ATTACHMENTS

5.1 Examples of publications

Examples of publications include:

- Monographs http://www.fs.fed.us/mntp/plan/index.htm
- Serials http://www.gpoaccess.gov/indicators/browse.html
- Journals http://www.ers.usda.gov/AmberWaves/
- Posters http://store.usgs.gov/historicmapsfromlca/images/LewisClarkPoster_p.pdf
- Maps http://www.epa.gov/wed/pages/ecoregions/tx%5Feco.htm and http://memory.loc.gov/cgi-bin/query/r?ammem/gmd:@field(NUMBER+@band(g7610+ct001267
- Application forms http://www.ed.gov/programs/jacobjavits/applicant.html
- Technical reports http://www.fs.fed.us/pnw/pubs/pnw_gtr621.pdf
- Handbook or manuals http://www.uscg.mil/ccs/cit/cim/directives/CIM/CIM_10360_3C.pdf
- ERIC Documents http://eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/2a/2f/df.pdf
- Juvenile activity and coloring books http://www.coastalscience.noaa.gov/education/ncbook.pdf
- Fact sheets http://www.epa.gov/safewater/lcrmr/lead.html and http://www.ojp.usdoj.gov/ovc/publications/factshts/ttac/fs000305.pdf
- Guides, travel brochures, and similar documents http://www.nps.gov/apco/
- USGS Open file reports http://pubs.usgs.gov/of/2005/1179/pdf/OFR-2005-1179.pdf
- Integrating resources http://www.irs.gov/irm/index.html and http://www.nationalatlas.gov/

5.2 Examples of published information not considered publications

Examples of published information that are not considered publications or whole publications are:

- Job vacancy notices or announcements
- Data input forms used to record information to be put into manual or computer record systems
- Forms that facilitate correspondence, such as memorandum or letterhead stock, envelopes, business cards, transmittal slips, and guidelines for correspondence performance.
- Personnel evaluation forms
- Solicitations for the awarding of procurements (these are not individual publications themselves but are published in a publication, similar to journal articles)
- Access passes or identification for automobiles, people or buildings
- Signs and bumper stickers that instruct
- Form letters designed to go to multiple recipients
- Agency control forms, handbooks, and manuals used in the management of property such as typewriters, paper, etc.

5.3 Publication terminology in English

Abstract
Academic dissertation
Adobe Acrobat Reader
Aeronautical chart
Almanac
Analysis
Annual Performance Plan
Annual Report
Appendices
Appendix
Atlas
Audit
Author
Authored
Authored by
Authors
Available in PDF
Bill
Biobibliography
Biography
Book
Book Illustration
Bookplate
Broadside
Budget
Bulletin
Calendar
Catalog
Chapter
Chart
Chronology
Clearinghouse
Collected Correspondence
Collected Works
Collections
Compendia
Compendium
Conference proceedings
Conference report
Congresses
Congressional Justification
Contract
Data warehouse
Database
Depository
Directory
Docs

Document
Documentaries
Edition
Electronic Journal
Encyclopedia
Environmental impact report
EIR
Environmental impact statement
EIS
Ephemera
Essay
Fact Sheet
Festschrift
For administrative use only
For internal use only
For official use only
For sale by the Superintendent of Documents
Form
Full-text
Gazetteer
Glossary
Grant
Guide
Guidebook
Handbook
Hearing
Impact statement
Index
Indices
Journal
Juvenile Literature
Laboratory Manual
Law
Legal Case
Legislation
Library
Manual
Manuscript
Map
Monograph
Nautical chart
News release
Newsletter
Notebook
Patent
PDF
Peer-reviewed journal
Performance report
Periodical
Pictorial Work

Plan
Popular Work
Poster
Price List
Print
Proceedings
Publication
Published
Published by
Pubs
Quarterly
Regulation
Regulatory
Report
Report number
Repository
Reprint
Reprinted
Request a hard copy
Resource
Resource Guide
Review
Review Literature
Revised
Sales
Scholarly journal
Scientific paper
Serial
Special volume
Statistical supplement
Statistic
Strategic plan
Study
Supplement
Survey
Table of contents
Table
Technical Report
Terminology
Theses
Thesis
Union List
Working paper
Workshop

5.4 Publication terminology in Spanish

Almacén de los datos
Almacén
Almanac
Análisis
Apéndice
Apéndices
Audiencia
Audiencias
Autor
Autores
Base de datos
Biblioteca
Biografía
Boletín
Boletín de noticias
Calendario
Cámara de compensación
Capítulo
Carta aeronáutica
Carta náutica
Cartas aeronáuticas
Cartas náuticas
Carteles
Casos Legales
Catálogo
Colecciones
Compendio
Con texto completo
Conferencia
Congresos
Contenido
Contrato
Cronología
Cuaderno
Depósito
Diccionarios geográficos
Directorio
Disertaciones Académicas
Disponible en PDF
Documento
Edición
Enciclopedia
Estadística
Extracto
Festschrift
Forma
Glosario

Guía
Indice
Informe
Informe Annual
Informe de la conferencia
Informe del sitio
Informe Técnico
Justificación del congreso
Legislación
Ley
Libro
Listas De la Unión
Listas De precios
Literatura Juvenil
Los diarios electrónicos
Manuale
Manuales De Laboratorio
Manuscritos
Mapas
Monografía
Narrativas Personales
Papel científico
Para el uso administrativo solamente
Para el uso interno solamente
Para el uso oficial solamente
Para la venta del superintendente de documentos
Patente
Periódico
Plan estratégico
Presupuesto
Publicación
Publicación Contraída
Publicado
Recurso
Regulación
Reimpreso
Revisado
Solicite una copia dura
Suplemento
Suplemento estadístico
Tabla
Terminología
Tesis
Trabajos Populares
Trimestral
Ventas
Volumen especial

5.5 File extensions in GPO's Catalog of U.S. Government Publications PURL server

Results of searches by the following file extensions in the U.S. Catalog of Government Publications (http://www.gpoaccess.gov/cgp/index.html) PURL server (http://purl.access.gpo.gov/maint/) on March 25, 2005.

File Extension	Number Found	Percentage	Notes
pdf	35360	73.8	34490 lower case, 870 capitalized
html	5293	11.05	5291 lower case, 2 capitalized
htm	5091	10.6	4954 lower case, 137 capitalized
txt	672	1.4	670 lower case, 2 capitalized
asp	624	1.3	All lower case
cfm	466	0.97	All lower case
shtml	106	0.22	All lower case
jsp	62		
shtm	53		
zip	49		
php	42		
exe	32		
mar	29		
aspx	22		
js	8		
avi	4		
wpd	3		
gif	3		
mov	3		
ppt	3		
sid	2		
xml	2		
hqx	1		
stm	1		
tif	1		

5.6 Other file extensions

Results of searches by these file extensions in the CGP PURL server on March 28, 2005.

File Extension	Number Found	Notes
aiff	0	
asf	0	
asmx	0	

au	172?	Most “au” not file extension
cif	Inconclusive results	
csv	0	
db	Inconclusive results	
dmg	0	
doc	220?	Most “doc” not file extension
dot	500?	Most “dot” not file extension but Dept. of Transportation acronym
eps	Inconclusive results	
fpt	0	
gz	0	
indd	0	
jar	0	
jfif	0	
kpg	0	
lit	Inconclusive results	
lwp	0	
m4a	0	
max	Inconclusive results	
mdb	Inconclusive results	
mdi	0	
mid	0	
midi	0	
mpu	Inconclusive results	
mpg	0	
moov	0	
ns2	Inconclusive results	
ns3	Inconclusive results	
ns4	Inconclusive results	
ocx	0	
p65	0	
pct	Inconclusive results	
pgm	0	
pl	Inconclusive results	
pmd	0	
pps	Inconclusive results	
ps	Inconclusive results	
psd	Inconclusive results	
pub	Inconclusive results	
qt	Inconclusive results	
ra	Inconclusive results	
ram	Inconclusive results	
rar	Inconclusive results	
rcd	0	

rm	Inconclusive results	
sea	Inconclusive results	
sit	Inconclusive results	
smi	Inconclusive results	
sql	0	
tga	0	
tmb	0	
uu	0	
uue	0	
wk1	0	
wma	0	
wmv	0	
wpt	0	
wpm	0	
z	Inconclusive results	
bmp	0	
class	0	
css	0	
dwg	0	
jpeg	0	
jpg	0	
mp3	0	
mp4	0	
mpeg	0	
mpg	0	
phtml	0	
png	0	
rtf	0	
swf	0	
tar	0	
wav	0	

Attachment #3: Blue Angel Rules

Two tests were applied by Blue Angel to determine if a document was in scope (i.e. considered to be an EPA publication). The first was to determine if a document was considered an in-scope publication, the second was to determine if the document is an EPA publication.

Rules to Determine if a Document is a Publication

These rules apply a test to see if a document is an in-scope publication. A document is considered to not be an in-scope publication if it meets any of the following criteria.

- **NP_UnsupportedType:** The document type is not supported.

- **NP_Abstract**: Indicates that a document is an abstract. This test checks if any of the following conditions are met:
    - The Subject or Title metadata field contains the word “abstract”
    - The Front Matter metadata field contains at least five (5) of the following words and phrases: “Abstract:”, “Citation:”, “Contact:”, “Division:”, “Branch:”, “Product Type:”, “Presented:”, “Related Entries:”

- **NP_Agenda:** Indicates that a document is a conference agenda. This test checks if all of the following conditions are met:
    - The Front Matter metadata field contains at least five (5) of the following words and phrases: “agenda”, “break”, “call to order”, “conference”, “cost”, “goal”, “goals”, “hotel”, “lodging”, “lunch”, “luncheon”, “master of ceremonies”, “meal”, “meals”, “opening comments”, “papers”, “presentation”, “presentations”, “registration”, “seminar”, “seminars”, “session”, “sessions”, “speaker”, “speakers”, “topic”, “track”, “travel information”, “welcome”, “workshop”
    - The First 250 Words, Subject, and Title metadata fields do not contain the word “proceedings”

- **NP_ConsentForm:** Indicates that a document is a consent form. This test checks if the Front Matter metadata field begins with the phrase "consent for"

- **NP_Docket:** Indicates that a document is a docket publication. This test checks if theDescription, Subject, or Title metadata field contains the phrase “Docket No.”

- **NP_Draft:** Indicates that a document is a draft. This test checks if any of the following conditions are met:
    - The Subject or Title metadata field begins or ends with the word “draft”
    - The First 250 Words metadata field contains the word “draft”

- **NP_Form:** Indicates that a document is a form. This test checks if the First 250 Words metadata field contains at least one of the following phrases: “amendment of solicitation”, “for instructions”, “For Sample Use Only”, “modification of contract”, “see instructions”, “type or print all information”

- **NP_Fragment:** Indicates that a document is a document fragment. This test checks if the Description or Subject metadata field contains at least one of the following phrases: “extracted page”, “extracted pages”, “from the”

- **NP_Instructions:** Indicates that a document is a set of form instructions. This test checks if the Front Matter metadata field begins with the phrase "instructions for"

- **NP_InternalSummaryMemo:** Indicates that a document is an internal summary memorandum. This test checks if the First 250 Words metadata field contains at least two of the following words and phrases: “Action:”, “Agency:”, “RFIP No.:”, “Summary:”, “Title:”

- **NP_Letter:** Indicates that a document is a letter. This test checks if any of the following conditions are met:
    - The First 250 Words metadata field contains the phrase "Dear <1-4 words>:" or “Dear <1-4 words>,” where <1-4 words> can be any set of one to four words.
    - The First 250 Words metadata field contains the phrase “letter from”

- **NP_MeetingAnnounce:** Indicates that a document is a meeting announcement. This test checks if all of the following conditions are met:
    - The Subject or Title metadata field contains the phrase “public meeting”
    - The First 250 Words, Subject, and Title metadata fields do not contain the word “proceedings”

- **NP_Memo:** Indicates that a document is a memorandum. This test checks if all of the following conditions are met:
    - The First 250 Words metadata field contains any of the following words: “memo”, “memorandum”
    - The First 250 Words metadata field contains at least two of the following words and phrases: “Attendees:”, “Date:”, “Date and Time:”, “From:”, “Location:”, “Re:”, “Subj:”, “Subject:”, “Time:”, “To:”
    - The First 250 Words, Subject, and Title metadata fields do not contain any of the following phrases: “memorandum of understanding”, “memorandum of agreement”, “memorandum of intent”

- **NP_MemoOfUnderstanding:** Indicates that a document is a memorandum of understanding. This test checks if the Subject or Title metadata field contains at least one of the following phrases: “memo of understanding”, “memorandum of agreement”, “memorandum of intent”, “memorandum of understanding”, “MOU”

- **NP_MetadataRecord:** Indicates that a document is a metadata record. This test checks if the Front Matter metadata field begins with the phrase "metadata record"

- **NP_Minutes:** Indicates that a document is a meeting minutes. This test checks if all of the following conditions are met:
    - The First 250 Words metadata field contains at least one of the following phrases: “Conference Call Summary”, “meeting minutes”, “meeting summary”, “public meeting”, “Stakeholders Meeting”, “summary meeting”, "summary minutes”
    - The First 250 Words, Subject, and Title metadata fields do not contain the word “proceedings”

- **NP_PurchaseOrder:** Indicates that a document is a purchase order. This test checks if the First 250 Words metadata field contains the phrase “purchase order”

- **NP_Readme:** Indicates that a document is a readme file. This test checks if the Front Matter metadata field begins with the word "readme"

- **NP_Solicitation:** Indicates that a document is a solicitation. This test checks if the First 250 Words metadata field contains at least one of the following phrases: “solicitation, offer, and award”, “this contract is a”, "type of solicitation”

- **NP_SOW:** Indicates that a document is a statement of work. This test checks if the First 250 Words metadata field contains the phrase “Statement of Work”

- **NP_SurveyForm:** Indicates that a document is a survey form. This test checks if any of the following conditions are met:
    - The Keywords metadata field contains at least one of the following words: "form", "questionnaire", "survey"
    - The First 250 Words metadata field contains the word “questionnaire”

- **NP_Testimony:** Indicates that a document is a testimony. Testimony is not considered a publication, as it would be included with the hearing and as such would be duplicative. This test checks if the Front Matter metadata field begins with any of the following words and phrases: “statement of”, “testimony”

Otherwise, a document is considered to be a Publication if it meets any of the following criteria:

- **QP_EnglLink1:** The Source Link Language is English and the Source Link Words imply that the document is a publication

- **QP_First250Words:** The First 250 Words imply that the document is a publication

- **QP_Funded:** The Front Matter contains the text along the lines of “publication was funded by grant or contract”

- **QP_ISBN:** An ISBN is found in the Front Matter or End Matter

- **QP_ISSN:** An ISSN is found in the Front Matter or End Matter

- **QP_SpanLink1:** The Source Link Language is Spanish and the Source Link Words imply that the document is a publication

Otherwise, the document is considered to not be a Publication.

Rules to Determine if a Publication is an EPA Publication

A document is deemed to be an EPA publication if all of the following criteria are not met:

- **QE_CFR:** The First 250 Words metadata field contains the string "CFR". Note that this is a string match and not a word or phrase match.

- **QE_CongressionalRecord:** The Subject or Title metadata field contains the phrase "congressional record"

- **QE_EPACDER:** The Front Matter or End Matter metadata field contains text referencing the EPA's Central Data Exchange Registration

- **QE_EPADirectory:** The Front Matter metadata field contains full text referencing the Environmental Protection Agency at least eight (8) times.

- **QE_FedReg:** The document is associated with the Federal Register

- **QE_NonUSEPA:** Indicates that a document is a non-U.S. E.P.A. publication. This test checks if any of the following conditions are met:
    - The Subject or Title metadata field contains any of the Agency or Abbreviation values found in Appendix: State Environmental Agencies
    - The Front Matter metadata field begins with any of the Agency or Abbreviation values found in Appendix: State Environmental Agencies

- **QE_TitleCFR:** The Title metadata field contains the word "CFR", the word "C.F.R.", or the phrase "Code of Federal Regulations"

- **QE_TitlePublicLaw:** The Title metadata field contains the phrase "Public Law"

AND any of the following criteria ***are*** met:

- **QE_AuthorEPA:** The Author metadata field contains full text referencing the Environmental Protection Agency

- **QE_DescriptionEPA:** The Description metadata field contains full text referencing the Environmental Protection Agency

- **QE_EndEPA:** The End Matter metadata field contains full text referencing the Environmental Protection Agency (see Algorithm EPATextFull).

- **QE_FrontEPA:** The Front Matter metadata field contains full text referencing the Environmental Protection Agency

- **QE_NonSeedEPA:** All of the following criteria are met:
    - The publication is not from a Seed URL
    - The Author, Description, Subject, or Title metadata field contains full text
    - referencing the Environmental Protection Agency, or abbreviated text referencing the EPA
    - The End Matter or Front Matter metadata field contains full text referencing the Environmental Protection Agency

- **QE_SubjectEPA:** The Subject metadata field contains full text referencing the Environmental Protection Agency

- **QE_TitleEPA:** The Title metadata field contains full text referencing the Environmental Protection Agency

Attachment #4: IIA Rules

The following table represents the final rules used by IIA. The “generalizable” column provides an indication of whether these rules can be generalized to other agencies. The “Score” column provides the weight assigned to each rule, and whether they were positive or negative rules (positive rules are indicators of in scope documents and negative rules are indicators of documents that are not in scope. The “Attribute” column denotes what attributes were examined by the harvester for each rule, and “Values” are the actual words or phrases that were examined.

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
1.2		EPA-hosted Federal Register Notices			
1.2.1	S		-3	object-title	Federal Register, "For Immediate release
1.2.2	S		-3	keyword	Federal Register,"For Immediate release"
1.2.3	S		-3	document-summary	Federal Register,"For Immediate release"
1.2.4	S		-9	theurl	fedrgstr
1.2.5	S		-3	epa_breadcrumbs	Federal Register
1.2.6	S		-3	links_and_labels	Federal Register
1.2.7	S		-3	headings	Federal Register Notice
1.2.6	S		-3	highlighted	Federal Register
1.3		EPA news releases			
1.3.1	S		3	document-text	@epa.gov
1.3.2	S		2	document-text	for immediate release
1.3.2	S		2	document-text	for immediate release
1.4		EPA approved content			
1.4.1	S		1	document-text	epa approved,"epa has approved"
1.4.2	S		3	links_and_labels	epa approved,"epa has approved"
1.4.3	S		3	object-title	epa approved,"epa-approved","epa has approved"
1.5		Letters			
1.5.1	Y		-3	document-text	dear,"sincerely","thank you"
1.6		Procurement Office			
1.6.1	Y		-3	document-text	Procurement Office

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
2		Official reports			
2.1	Y	PDF reports			
2.1.1		PDF	+	Object-type	application/pdf
2.1.2					
2.1.2.1	S		2	document-text	fact sheet, "copies of this report available from", "copies of this fact sheet available from","List of Tables", "List of Images", "Table of Contents", "Environmental Impact Statement", EIS, "Environmental Impact report", EIR, "Request a hard copy", "Resource guide", "Technical Report", "Working paper", "Review Literature", "intentionally left blank"
2.1.2.2	Y		1	document-text	report,contents,introduction,references,revised
2.1.3	S		3	document-summary	Final report, "Fact Sheet", "Environmental Impact Statement", Proceedings
2.1.4	S		3	object-title	epa order,"Final report", "Fact Sheet", "Environmental Impact Statement", Proceedings
2.1.6	Y		-5	object-title	Draft
2.1.7	Y		-3	referrer_url	draft
2.1.8	S		3	Webi_description	Final report, "Fact Sheet", "Environmental Impact Statement", Proceedings
2.1.9	S		3	Webi_title	Final report, "Fact Sheet", "Environmental Impact Statement", Proceedings
2.1.10	Y		-3	Webi_title	Draft
2.1.11	Y		1	highlighted	fact sheet
2.1.12	S		1	links_and_labels	epa order
2.1.13		Child pages fact sheet			
2.1.13.1	Y		1	document-text	fact sheet
2.1.13.2	Y		1	object-title	fact sheet
2.1.13.3	Y		1	object-title	fact sheet

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
2.2		HTML reports			
2.2.1	Y	HTML	+	Object-type	text/html
2.2.2					
2.2.2.1	S		2	document-text	fact sheet, "copies of this report available from", "copies of this fact sheet available from", "List of Tables", "List of Images", "Table of Contents", "Environmental Impact Statement", EIS, "Environmental Impact report", EIR, "Request a hard copy", "Resource guide", "Working paper", "Review Literature", "Study Purpose", "Funding organization","intentionally left blank"
2.2.2.2	Y		1	document-text	report,contents,Introduction,references,revised
2.2.3	S		3	document-summary	Report, "fact sheet", "copies of this report available from", "copies of this fact sheet available from", Introduction, Content, References, "List of Tables", "List of Images", Attachments, "Table of Contents", "Environmental Impact Statement", EIS, "Environmental Impact report", EIR, "Proceedings of", "Request a hard copy", "Resource guide", "Working paper", Revised, "Review Literature", "Study Purpose", "Funding organization", "Funding provided by",
2.2.4	S		3	Keyword	fact sheet, "copies of this report available from", "copies of this fact sheet available from", Introduction, Content, References, "List of Tables", "List of Images", Attachments, "Table of Contents", "Environmental Impact Statement", "Environmental Impact report", EIR, "Proceedings of", "Request a hard copy", Resource guide", "Working paper", Revised, "Review Literature", "Study Purpose", "Funding organization", "Funding provided by",

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
2.2.11	S		3	Webi_description	Report, "fact sheet", "copies of this report available from", "copies of this fact sheet available from", Introduction, Content, References, "List of Tables", "List of Images", Attachments, "Table of Contents", "Environmental Impact Statement", EIS, "Environmental Impact report", "Proceedings of", "Request a hard copy", Resource guide", "Technical Report", "Working paper", Revised, "Review Literature", "Study Purpose", "Funding organization", "Funding provided by",
2.2.12	S		-3	Webi_title	Draft
2.2.13	Y		1	highlighted	fact sheet
2.1.14	N		1	links_and_labels	epa order
2.1.15			3	object-title	epa order,"Final report", "Fact Sheet", "Environmental Impact Statement", Proceedings
2.2.16	S	Child pages fact sheet			
2.2.16.1			1	document-text	fact sheet
2.2.16.2	Y		1	object-title	fact sheet
2.2.16.3	Y		1	object-title	fact sheet

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
3		EPA Posters			
		EPA Posters			
3.1		Descriptive rules			
3.1.1	Y		3	Object-type	image,media-video
3.1.2	Y		3	referrer_url	poster
3.1.3	Y		3	img_alt	Poster
3.1.4	Y		2	Highlighted	poster
4		EPA Program Descriptions			
		EPA Program Descriptions			
4.1		Keyword			
4.1.1	Y		2	document-text	official business,"program report"
4.1.2	Y		3	object-title	official business,"program report","program update"
4.1.3	Y		3	links_and_labels	official business,"program report","program update"
4.1.4	Y		2	Metadata	Geographic Area, "Project Officer"
4.1.5			3	Webi_title	official business,"program report","program update"
5		EPA Publications			
5.1		EPA Publications listed at NEPIS			
5.1.1	N		3	referrer_url	nepis.epa.gov/pubtitle
5.2		Publications by EPA researchers			
5.2.1	S		2	document-text	Source Document,"Agency Work Group Review", "Verification Date", "EPA Contacts", "Supporting Studies", "Quantitative Estimate", "EPA Documentation"

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
6		Agency Orgcharts			
6.1		Agency Orgcharts Features			
6.1.1	Y		3	Object-type	text/html, application-acrobat-pdf
6.1.2	Y		3	object-title	Organization* Chart, Organization
6.1.3	Y		3	document-summary	Organization* Chart, Organization, "Office of",
6.1.4	Y		2	Keyword	Organization* Chart, Organization. "Office of",
6.1.5	Y		2	Headings	Organization* Chart, Organization. "Office of",
6.1.6	Y		3	links_and_labels	Organization Chart, Organization, "Office of",
6.1.7	Y		3	img_alt	Organization Chart, Organization, "Office of",
6.1.7	Y		3	img_alt	Organization Chart, Organization, "Office of",
6.1.8	Y		3	Webi_keywords	Organization Chart, Organization
6.1.9	Y		1	Webi_description	Organization* Chart, Organization
6.2		Known related sources			
6.1.7			1	img_alt	Organization Chart, Organization, "Office of",
7		Agency Press Releases			
7.1		Indicators			
7.1.1	Y		1	document-text	Press Release
7.1.2	Y		3	object-title	Press Release
7.1.3	Y		3	labels	Press Release
7.1.4	Y		3	Keyword	Press Release
7.1.5	Y		3	links_and_labels	News Releases feed, "in the news"
7.1.6	Y		3	Webi_keywords	Press Release
7.1.7	Y		3	Webi_title	Press Release

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
7.2		Known agency subsystems containing Press Releases and Related Publications			
7.2.1	N		3	theurl	gov/newsroom/newsreleases ,opa/admpress
7.3		Public Service Announcements			
7.3.1			3	document-text	Public Service Announcement, PSA
7.3.2	Y		3	object-title	Public Service Announcement, PSA
7.3.3	Y		3	links_and_labels	Public Service Announcement, PSA
7.3.4	Y		2	Object-type	application-audio-mp3, text/html, application-adobe-pdf
7.3.5	N		3	referrer_url	gov/emergenc/katrina/outreach
7.3.6	Y		3	Webi_title	Public Service Announcement, PSA
8		Agency Advisories and Bulletins			
8.1		Advisories Indicators			
8.1.1	Y		3	object-title	Advisory on, "Advisory by"
8.1.2	Y		3	links_and_labels	Advisory on, "Advisory by"
8.1.3	Y		2	Headings	Advisory on, "Advisory by"
8.1.4	Y		1	document-text	Availability, Committee, Chair, "Table of contents", Abstract, "charge to subcommittee", "response by subcommittee"
8.1.5	Y		3	Webi_title	Advisory on, "Advisory by"

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
9		Funding Opportunities Announcements			
9.1		Funding Op. Indicators			
9.1.1	S		2	document-text	Solicitation, "Opening Date", "Closing Date", Eligibility, Submissions, "Application Form*", "Synopsis of Program", "funding opportunity", "award information", "under a grant", "federal grant","cooperative agreement", "form 424"
9.1.2	S		2	Headings	Solicitation, "Opening Date", "Closing Date", Eligibility, "Technical Contact", Submissions, "Application Form*", "Synopsis of Program", "funding opportunity", "award information", "under a grant", "federal grant","cooperative agreement", "standard form 424"
10		Environmental Indicators Documents			
11		Datasets and models			
11.1		Dataset indicators			
11.1.1	Y		1	Object-type	application-executable, media-archive, text/text
11.1.2	S		3	EIMS_Information_Type	Model,Dataset
12		EPA official responses to public comments			
12.1		ADI Control numbers			
12.1.1	S		3	ADI_Control_Number	*

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
13		Official Directives			
13.1	Y		3	object-title	guidance,"reporting guide"
14 pivot=		Parent-child rules			
14.1		Parent pages			
14.1.1	Y		3	links_and_labels	Chapter (\d+),"Section (\d+)","Appendix","Introduction","Cover","Findings"
14.2		Leaf pages			
14.2.1	Y		-3	headings	Chapter (\d+),"Section (\d+)"
14.2.2	Y		-3	object-title	Chapter (\d+),"Section (\d+)"
14.2.3	Y		-3	highlighted	Chapter (\d+),"Section (\d+)"
14.2.4	Y		-3	headings	Section (\d+),"Section (\d+)"
14.2.5	Y		-3	object-title	Section (\d+),"Section (\d+)"
14.2.6	Y		-3	highlighted	Section (\d+),"Section (\d+)"
14.2.7	Y		-3	headings	Introduction,"Cover","Findings"
14.2.8	Y		-3	object-title	Introduction,"Cover","Findings"
14.2.9	Y		-3	highlighted	Introduction,"Cover","Findings"
14.2.10	Y		-3	headings	Introduction,"Cover","Findings"
14.2.11	Y		-3	object-title	Introduction,"Cover","Findings"
14.2.12	Y		-3	highlighted	Introduction,"Cover","Findings"
14.3		Links or text			
14.3.1	Y	Mostly Links	3	@mostlylinks	
14.3.2	Y	Mostly Text	-3	@mostlytext	
15		Supplemental rules			
15.1	Y	Mostly Links	-6	@links	
15.2	N		3	theurl	airtrends/aqtrnd96/general
15.3	N		20	theurl	ttn/naaqs/ozone/areas
15.4	N		20	theurl	airmarket/emissions/raw/data
15.5	Y		-6	headings	permit
15.6	Y		-6	theurl	permit

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
15.6	Y		-6	object-title	permit
15.6	Y		-6	object-title	permit
21		Special handling			
21.1		Copyrighted material - special handling			
21.1.1	Y		3	document-text	copyright, "All rights reserved","Authorized use only","Not for distribution"
21.1.2	Y		3	copyright	*
21.2		Internal agency - special handling			
21.2.1	Y		3	headings	For official use,"For internal use","For administrative use","For operational use"
21.2.2	Y		2	document-text	For official use,"For internal use","For administrative use","For operational use"
21.3		Classified or restricted - special handling			
21.3.1	Y		3	headings	Classified,restricted
21.4		Documents with disclaimers			
21.4.1	Y		3	links_and_labels	Disclaimer
21.4.2	Y		2	Highlighted	Disclaimer
21.4.3	Y		1	document-text	Disclaimer
22		Works in progress			
22.1	Y		3	document-text	this is a test,"do not publish","limited distribution","not for distribution"
22.2	Y		3	theurl	text

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
30		System-generated rules			
30.1		Rules applied to problem docment			
30.1.2			1	epa_breadcrumbs	water
30.1.3			1	epa_breadcrumbs	great lakes
30.1.4			1	epa_breadcrumbs	publications
30.1.5			1	epa_breadcrumbs	document
30.1.6			1	epa_breadcrumbs	system
30.1.7			1	epa_breadcrumbs	pollution
30.1.8			1	epa_breadcrumbs	environmental publications
30.1.9			1	epa_breadcrumbs	nepis
30.1.10			1	epa_breadcrumbs	toxics strategy
30.1.11			1	epa_breadcrumbs	science
30.1.12			1	theurl	ord/webpubs
30.1.13			1	theurl	projsum
30.1.14			1	theurl	safewater
30.1.15			1	tssms	download
30.1.16			1	tssms	safewater
30.1.17			1	epa_breadcrumbs	information
30.1.18			1	tssms	Clariton
30.1.19			1	object-title	epa (\d+)
30.1.20			1	subject	innovative hazardous waste
30.1.21			1	theurl	criteria
30.1.22			1	author	office of water
30.1.23			1	theurl	swertio1
30.1.24			1	object-title	drinking
30.1.26			-1	labels	(\d+)
30.1.27			-1	object-title	region
30.1.30			-1	referrer_url	yosemite.epa.gov/r10

RuleID	Description	Score	Attribute	Values
30.1.31		-1	referrer_url	air
30.1.32		-1	links_and_labels	(\d+)
30.1.35		-1	links_and_labels	yosemite.epa.gov/r10
30.1.37		-1	links_and_labels	naaqs ozone areas plant
30.1.43		-1	referrer_url	(\d+)
30.1.47		-1	document-summary	age a gt
30.1.56		-1	object-title	ets cem
30.1.60		-1	referrer_url	ttp www.epa.gov/enviro.html
30.1.70		-1	document-summary	chemicals
30.1.71		-1	epa_contacts	415 (\d+)
30.1.74		-1	highlighted	co (\d+)
30.1.75		-1	labels	station (\d+)
30.1.77		-1	object-title	station unit
30.1.78		-1	referrer_url	raw data
30.1.79		-1	theurl	raw/data
30.2	Documents linking to problem docment			
30.2.3		1	object-title	ttn
30.2.4		1	webi_keywords	technology
30.2.10		-1	object-title	draft report
30.2.11		-1	theurl	region5/water
30.2.12		-1	theurl	oust
30.2.13		-1	theurl	uic
30.2.14		-1	webi_keywords	underground storage
30.2.15		-1	webi_keywords	agency grants
30.2.16		-1	webi_title	jobs through recycling
30.2.19		-1	tssms	indicate
30.2.20		-1	tssms	werust1
30.2.21		-1	theurl	glrppr.org/hubs
30.2.22		-1	tssms	eg5oh2o
30.2.24		-1	theurl	fedlaws
30.2.25		-1	theurl	workshop_slides

RuleID	Description	Score	Attribute	Values
30.2.26		-1	theurl	presentations
30.2.27		-1	theurl	envindicators/roe
30.2.28		-1	theurl	water/uic/presentations
30.3	Documents linked fromproblem docment			
30.3.3		1	object-title	ttn
30.3.4		1	webi_keywords	technology
30.3.6		-1	img_alt	disclaimer
30.3.7		-1	webi_description	state
30.3.9		-1	tssms	eg5oopa
30.3.10		-1	object-title	draft report
30.3.11		-1	theurl	region5/water
30.3.12		-1	theurl	oust
30.3.13		-1	theurl	uic
30.3.14		-1	webi_keywords	underground storage
30.3.15		-1	webi_keywords	agency grants
30.3.16		-1	webi_title	jobs through recycling
30.3.18		-1	links	tribal
30.3.19		-1	tssms	indicate
30.3.20		-1	tssms	werust1
30.3.21		-1	theurl	www.glrppr.org/hubs
30.3.22		-1	tssms	eg5oh2o
30.3.23		-1	tssms	paoswer
30.3.24		-1	theurl	fedlaws
30.3.25		-1	theurl	workshop_slides
30.3.26		-1	theurl	presentations
30.3.27		-1	theurl	www.epa.gov/envindicators/roe
30.3.28		-1	theurl	water/uic/presentations

RuleID	Description	Score	Attribute	Values
60	System generated rules second crawl			
60.1	Rules applied to problem docment			
60.1.1		2	contact_url	ttn/naaqs/ozone/contactus
60.1.2		2	object-title	epa ttn naaqs
60.1.3		2	document-text	nox co so2
60.1.4		2	document-text	emission home page
60.1.5		2	document-text	epa home privacy
60.1.6		2	document-text	transport of ozone
60.1.7		2	document-text	resources file utilities
60.1.8		2	document-text	page ozone implementation
60.1.9		2	links_and_labels	home page
60.1.10		2	object-title	ttn naaqs
60.1.11		2	document-text	drinking water
60.1.12		2	document-text	scc descriptions
60.1.13		-2	document-text	to prairies
60.1.14		2	document-summary	high-quality scientific
60.1.15		2	contact_url	maia/html/comments
60.1.16		2	labels	sheet
60.1.17		2	webi_title	sheet
60.1.18		2	webi_keywords	sheet
60.1.19		2	object-title	water
60.1.20		2	headings	sheet
60.1.21		2	author	of
60.1.22		2	object-title	water
60.1.23		2	highlighted	totals
60.1.24		2	webi_keywords	brownfields
60.1.25		2	links	and air quality

RuleID		Description	Score	Attribute
60.1.26		2	epa_breadcrumbs	emission trends data
60.1.27		2	links_and_labels	http://www.epa.gov/ttn/naaqs/ozone/areas/index.htm
60.1.28		2	links_and_labels	home page
60.1.29		2	object-title	sheet
60.1.30		-2	object-title	sector resources
60.1.31		-2	object-title	ncee publications regulatory
60.1.32		-2	keyword	tsca valuation value
60.1.33		-2	keyword	legislation market
60.1.34		-2	keyword	control cost cba
60.1.35		-2	document-summary	records are classified
60.1.36	Y	-2	document-summary	office in charge
60.1.37	Y	-2	document-summary	analyses of policies
60.1.38		-2	document-summary	related to epa's
60.1.39	Y	-2	object-title	powerpoint
60.1.40		-2	labels	perfect
60.1.43		-2	keyword	tradeoff trading tsca
60.1.44		-2	keyword	regulation regulatory release
60.1.45		-2	keyword	producer program
60.1.46		-2	keyword	permit
60.1.47		-2	keyword	permit pesticide policies
60.1.48		-2	keyword	omb permit pesticide
60.1.49		-2	keyword	occupational omb permit
60.1.50		-2	keyword	natural occupational omb
60.1.51		-2	keyword	hazardous health human
60.1.52		-2	keyword	equity estimation evaluation
60.1.53		-2	document-summary	protection agency's
60.1.54		-2	document-summary	prevention roundtable
60.1.57		-2	referrer_url	//cfpub.epa.gov/clearinghouse/index.cfm?topicid=c10
60.1.58		-2	links	economic analyses

RuleID		Description	Score		Attribute	Values
60.1.59	Y			-2	author	printing office
60.1.60				-2	subject	pages
60.1.61				-2	referrer_url	//www.epa.gov/imr/download/user/
60.1.62				-2	object-title	register
60.1.64				-2	document-summary	assistance
60.2		Documents linking to problem docment				
60.2.1				2	document-text	foia grants/procurement laboratory
60.2.2				2	document-text	agriculture brownfields cleanup
60.2.3				2	document-text	topics regional adminstrator
60.2.4				2	document-text	and workshops maps
60.2.5				2	document-text	amp development u.s
60.2.6				2	document-summary	and assessment initiative
60.2.7				-2	document-text	public notices
60.2.8				-2	contact_url	region5/water/r5water_comments
60.2.9				-2	document-text	injection control regulations
60.2.10				-2	links_and_labels	notices announcements
60.2.11				-2	document-text	topics other local
60.2.12				-2	document-text	landscape ecology environmental
60.2.13				-2	document-text	satisfaction survey uic
60.2.15				-2	document-text	financing business assistance
60.2.16				-2	document-text	through recycling
60.2.17				-2	document-text	facilities mines_count mines
60.2.19	Y			-2	document-text	hub
60.2.20	Y			-2	document-text	by keyword table
60.3		Documents linked fromproblem docment				
60.3.1				2	document-text	office of wetlands
60.3.2	Y			2	document-text	since the original publication

RuleID	Description	Score	Attribute	Values
60.3.3		2	document-text	is entirely drawn
60.3.4		2	webi_keywords	training and certification
60.3.5		2	webi_keywords	of watershed training
60.3.6		2	document-text	research amp development
60.3.7		2	document-text	gt icr gt
60.3.8		-2	document-text	prevention resource exchange
60.3.9		2	headings	homepage epa home
60.3.10		2	webi_title	of watershed training
60.3.11		2	webi_keywords	of watershed training
60.3.12		2	document-summary	scientific information on
60.3.14		2	subject	emission inventory conference
60.3.15		2	webi_keywords	and certification/ ecosystems
60.3.16		2	img_alt	of watershed training
60.3.17		2	links_and_labels	www.epa.gov/ow/search.html
60.3.19		2	referrer_url	www.epa.gov/ttn/chief/conference/ei13/index.html
60.3.20		2	referrer_url	www.epa.gov/ne/npdes/mirantkendall/index.html
60.3.22		2	webi_keywords	training and certification/
60.3.23		2	object-title	envirofacts warehouse icr
60.3.24		2	contact_url	enviro/html/ef_feedback
60.3.25		2	webi_keywords	education training
60.3.26		2	object-title	envirofacts warehouse
60.3.27		-2	document-summary	great lakes regional
60.3.28		-2	img_alt	pollution prevention roundtable
60.3.29		-2	object-title	management for schools
60.3.30		-2	theurl	hubs/keyword_search