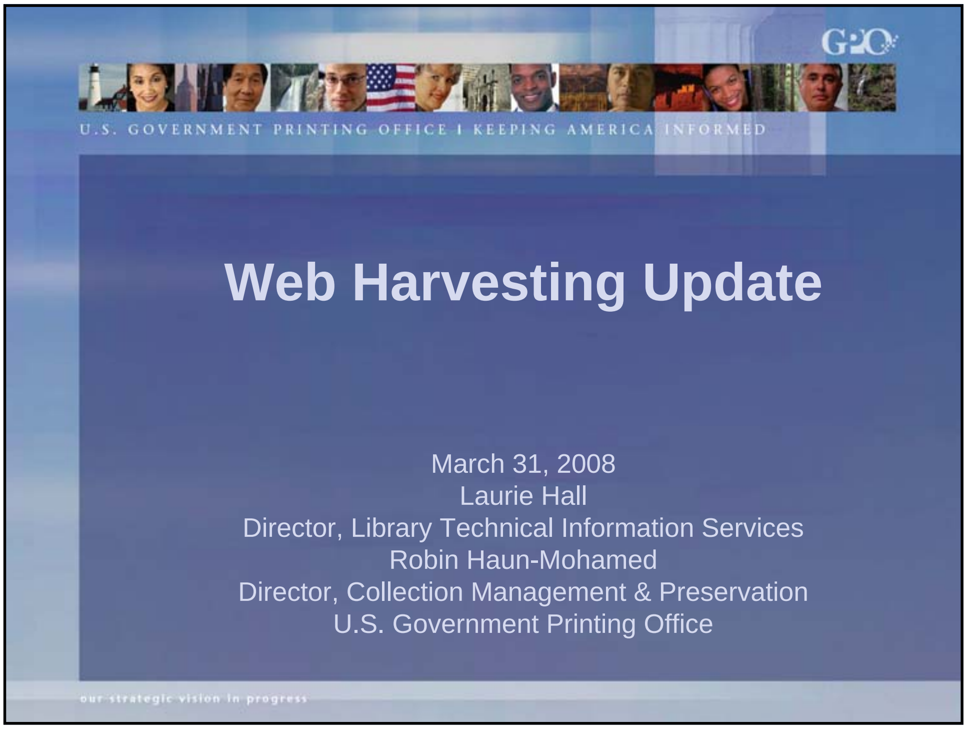
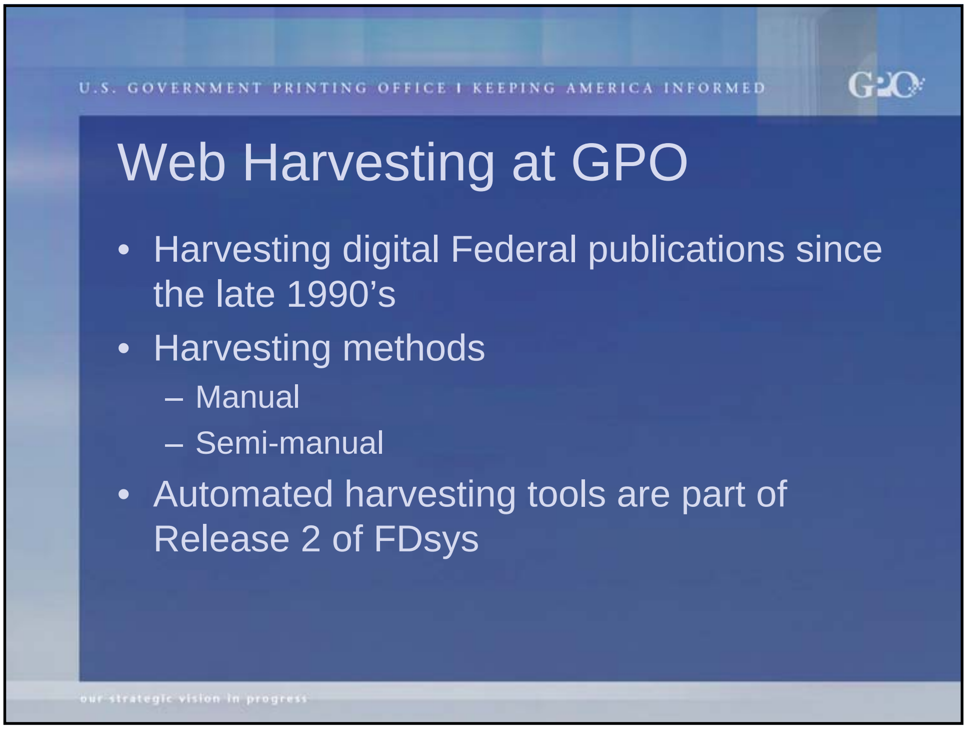


U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

Web Harvesting Update

March 31, 2008
Laurie Hall
Director, Library Technical Information Services
Robin Haun-Mohamed
Director, Collection Management & Preservation
U.S. Government Printing Office

our strategic vision in progress

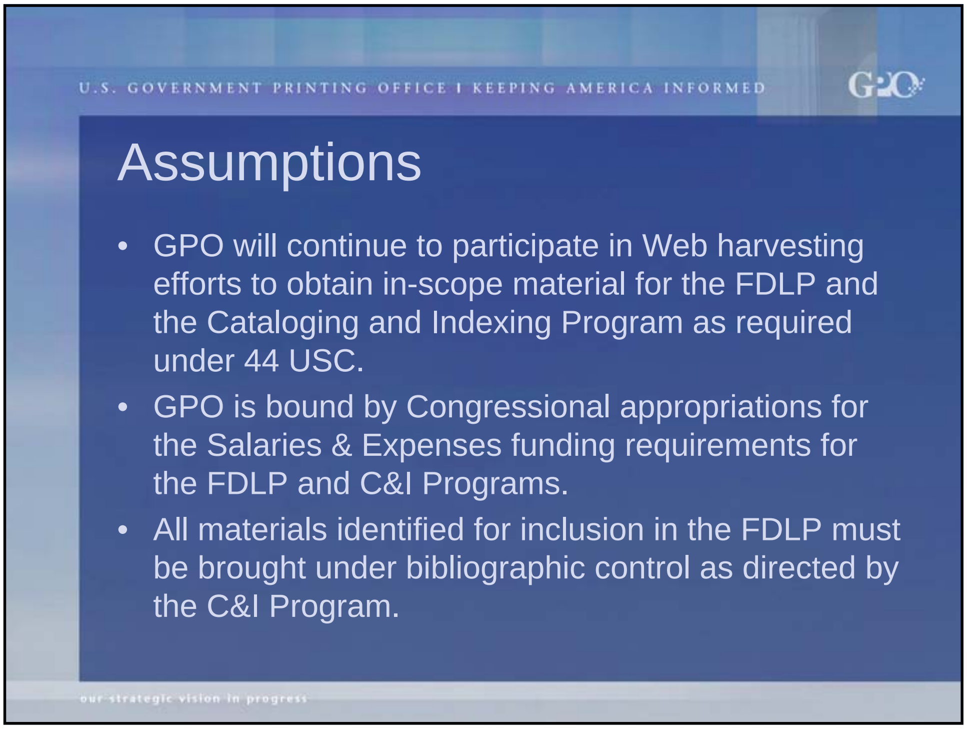


U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

Web Harvesting at GPO

- Harvesting digital Federal publications since the late 1990’s
- Harvesting methods
  - Manual
  - Semi-manual
- Automated harvesting tools are part of Release 2 of FDsys

our strategic vision in progress

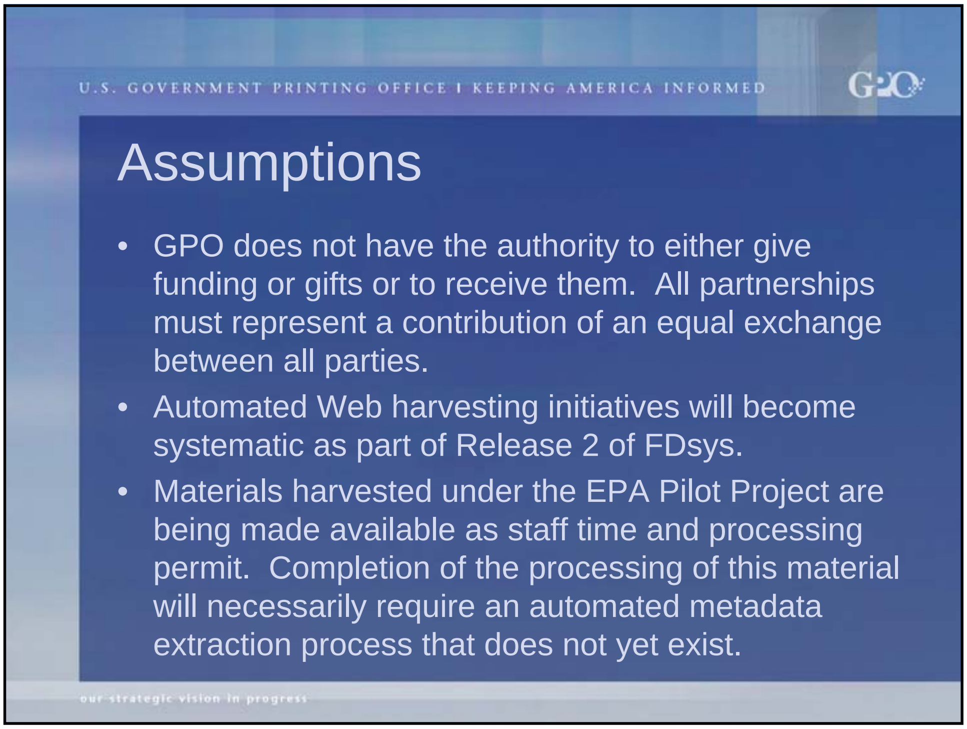


U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

Assumptions

- GPO will continue to participate in Web harvesting efforts to obtain in-scope material for the FDLP and the Cataloging and Indexing Program as required under 44 USC.
- GPO is bound by Congressional appropriations for the Salaries & Expenses funding requirements for the FDLP and C&I Programs.
- All materials identified for inclusion in the FDLP must be brought under bibliographic control as directed by the C&I Program.

our strategic vision in progress

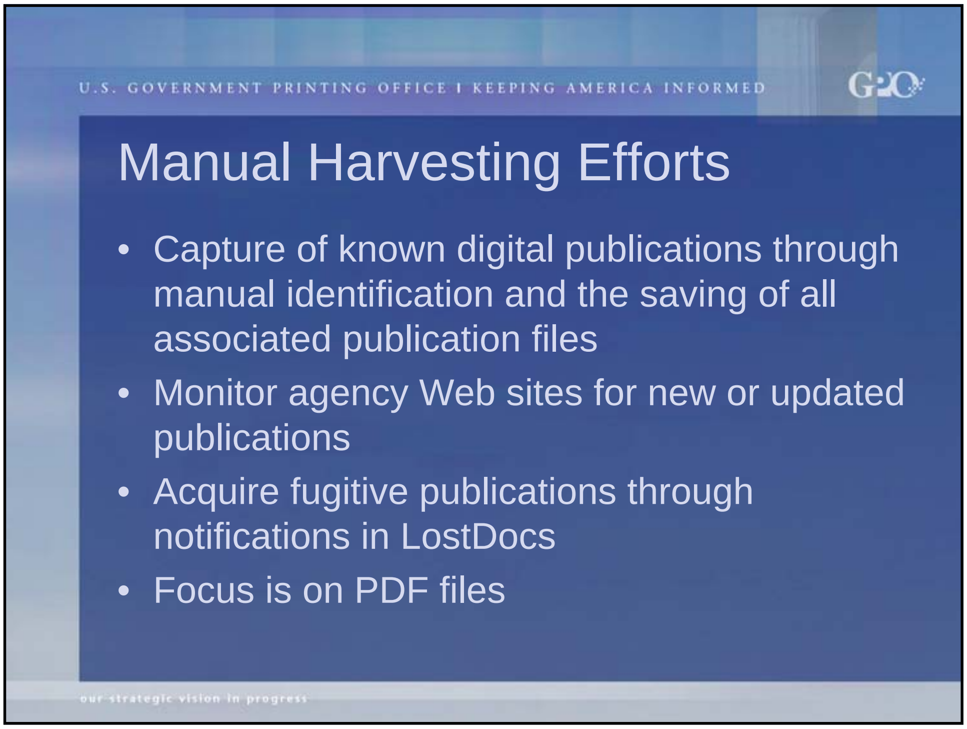


U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

Assumptions

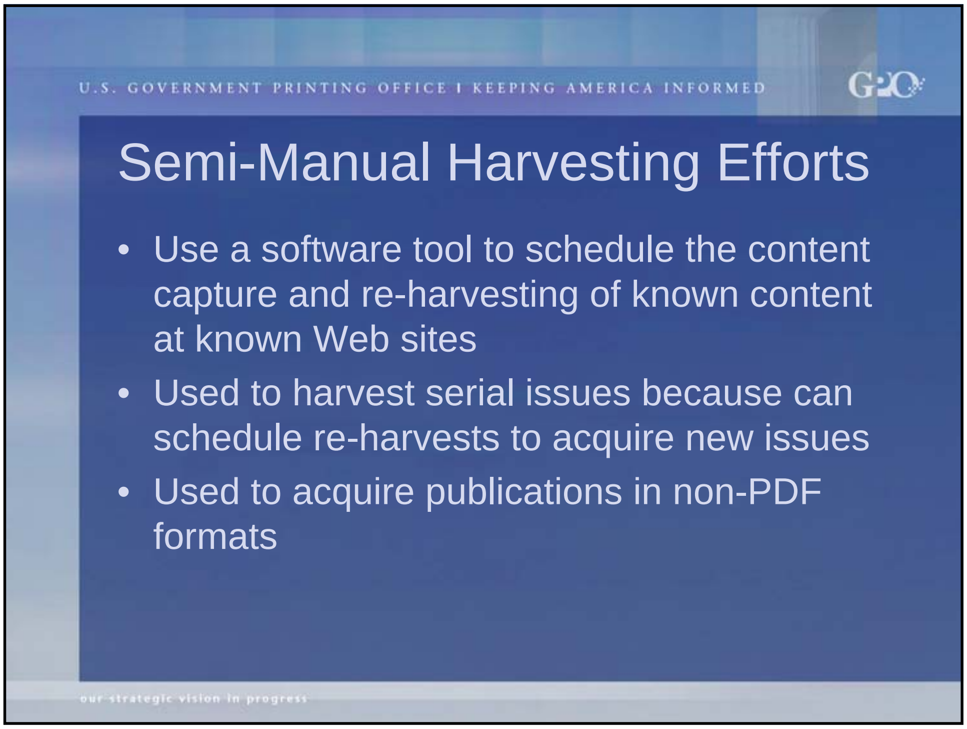
- GPO does not have the authority to either give funding or gifts or to receive them. All partnerships must represent a contribution of an equal exchange between all parties.
- Automated Web harvesting initiatives will become systematic as part of Release 2 of FDsys.
- Materials harvested under the EPA Pilot Project are being made available as staff time and processing permit. Completion of the processing of this material will necessarily require an automated metadata extraction process that does not yet exist.

our strategic vision in progress



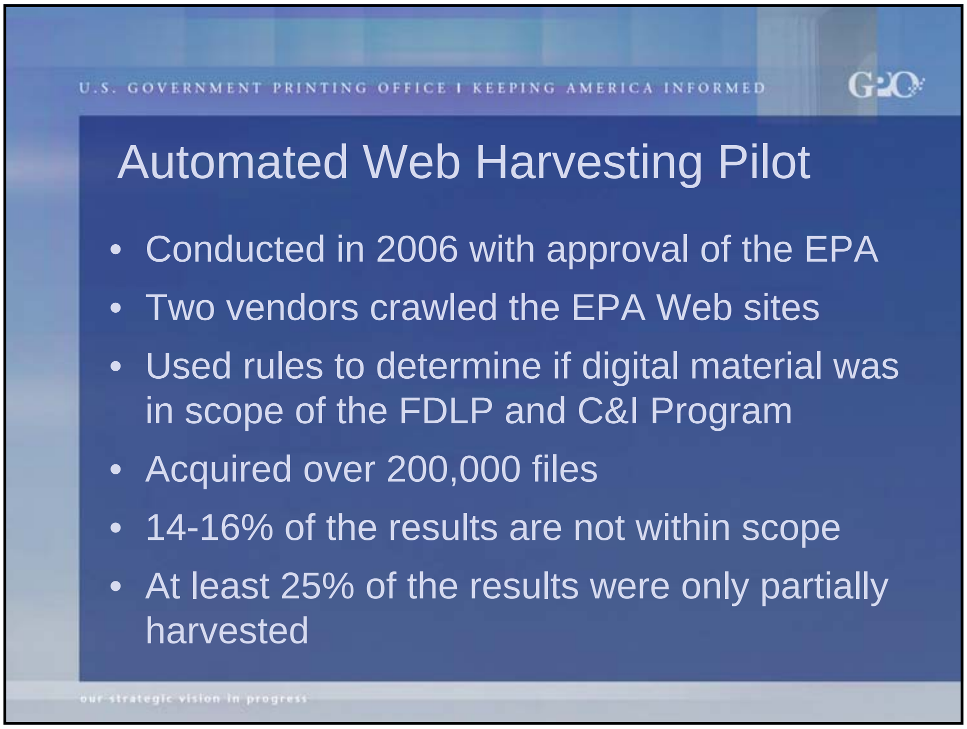
Manual Harvesting Efforts

- Capture of known digital publications through manual identification and the saving of all associated publication files
- Monitor agency Web sites for new or updated publications
- Acquire fugitive publications through notifications in LostDocs
- Focus is on PDF files



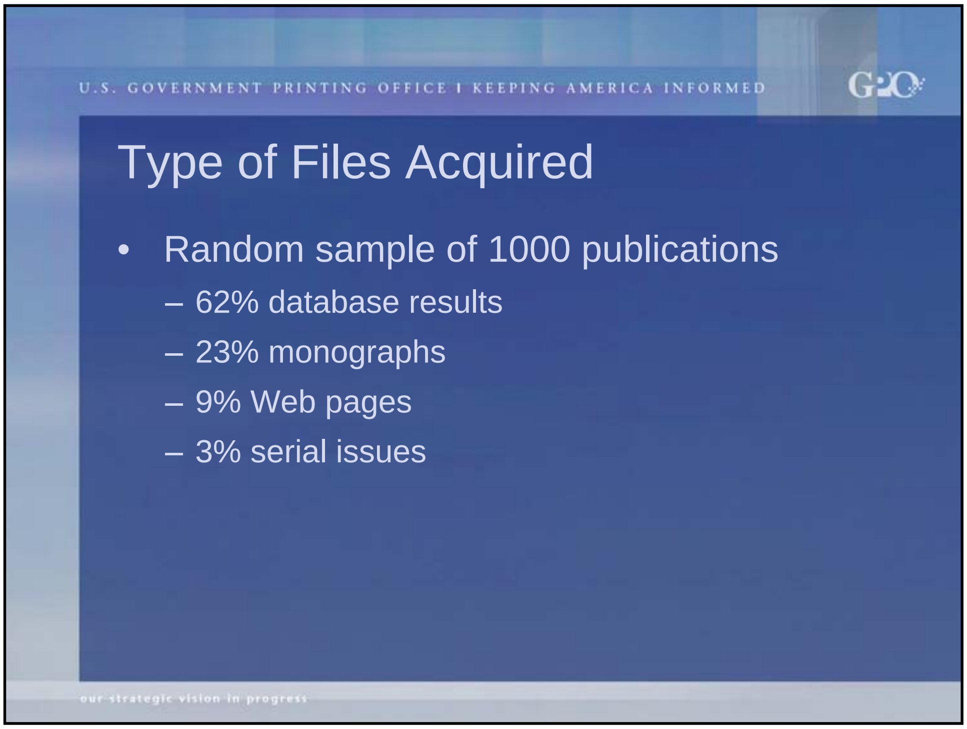
Semi-Manual Harvesting Efforts

- Use a software tool to schedule the content capture and re-harvesting of known content at known Web sites
- Used to harvest serial issues because can schedule re-harvests to acquire new issues
- Used to acquire publications in non-PDF formats



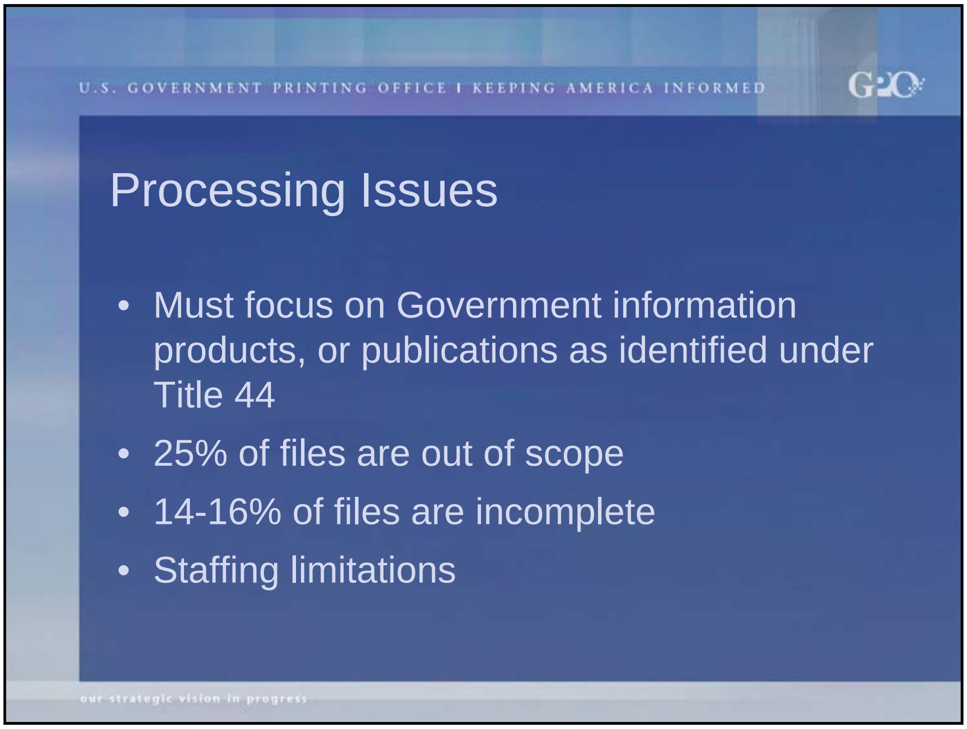
Automated Web Harvesting Pilot

- Conducted in 2006 with approval of the EPA
- Two vendors crawled the EPA Web sites
- Used rules to determine if digital material was in scope of the FDLP and C&I Program
- Acquired over 200,000 files
- 14-16% of the results are not within scope
- At least 25% of the results were only partially harvested



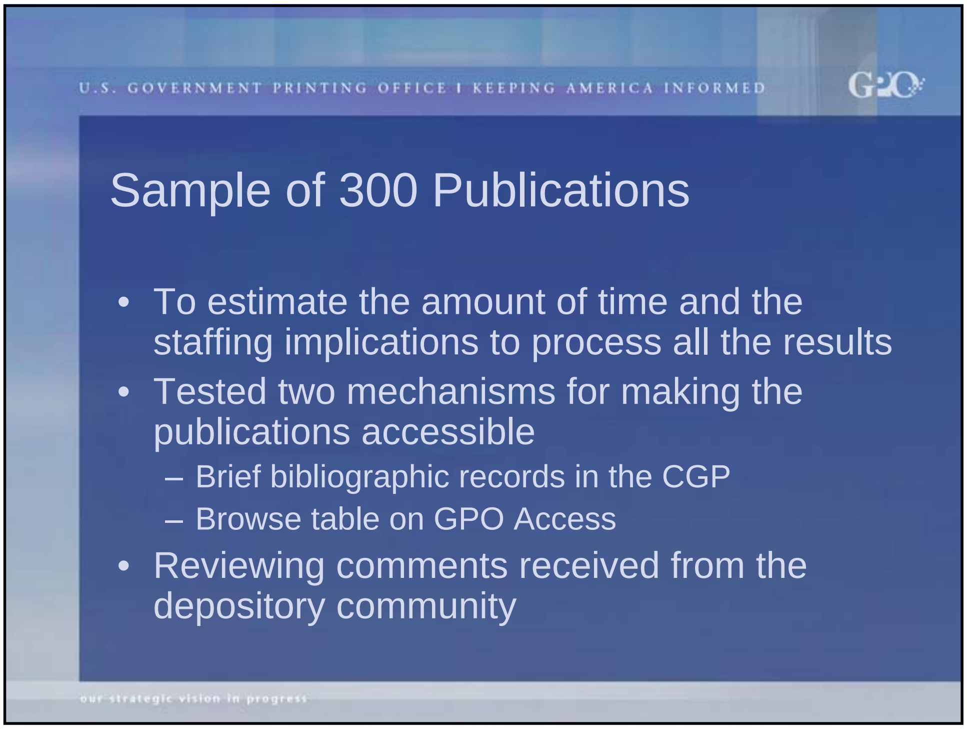
Type of Files Acquired

- Random sample of 1000 publications
  - 62% database results
  - 23% monographs
  - 9% Web pages
  - 3% serial issues



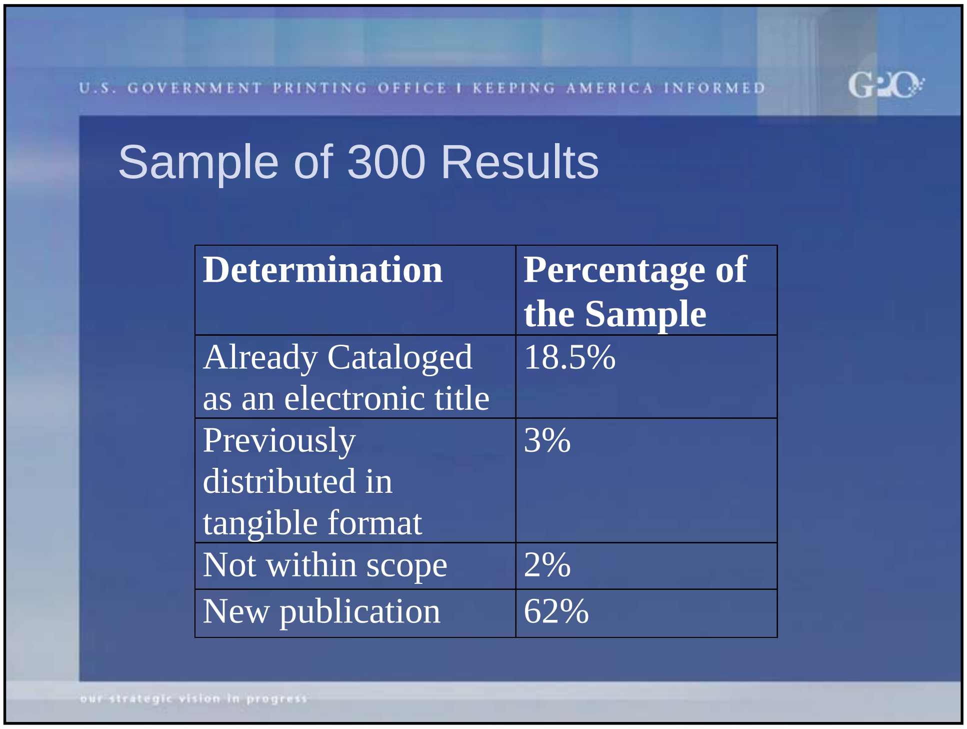
Processing Issues

- Must focus on Government information products, or publications as identified under Title 44
- 25% of files are out of scope
- 14-16% of files are incomplete
- Staffing limitations



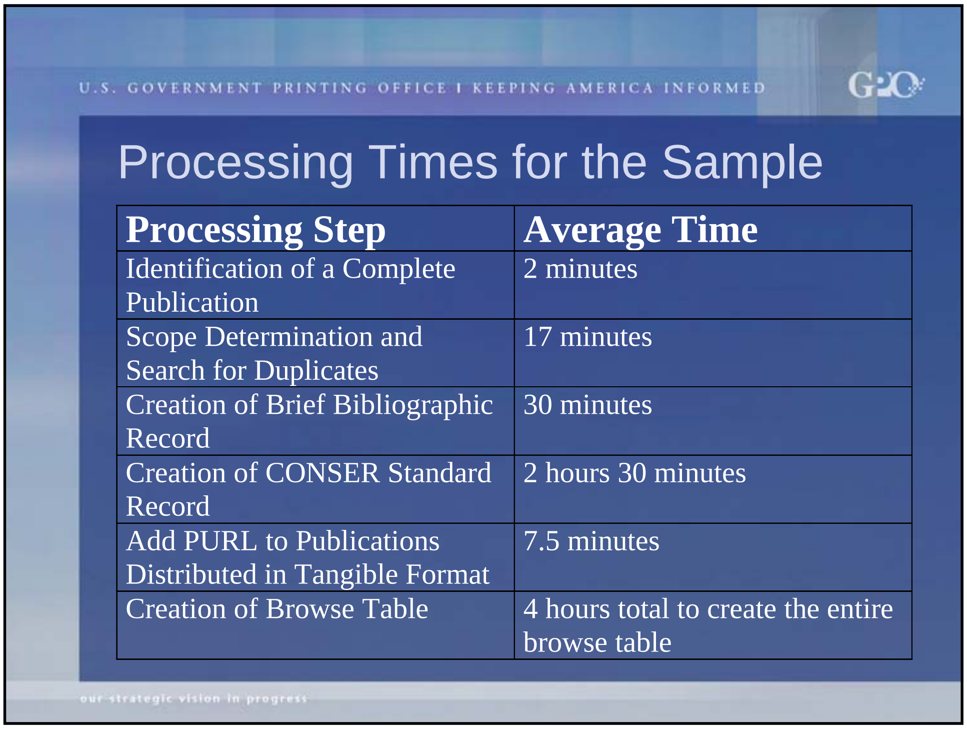
Sample of 300 Publications

- To estimate the amount of time and the staffing implications to process all the results
- Tested two mechanisms for making the publications accessible
  - Brief bibliographic records in the CGP
  - Browse table on GPO Access
- Reviewing comments received from the depository community



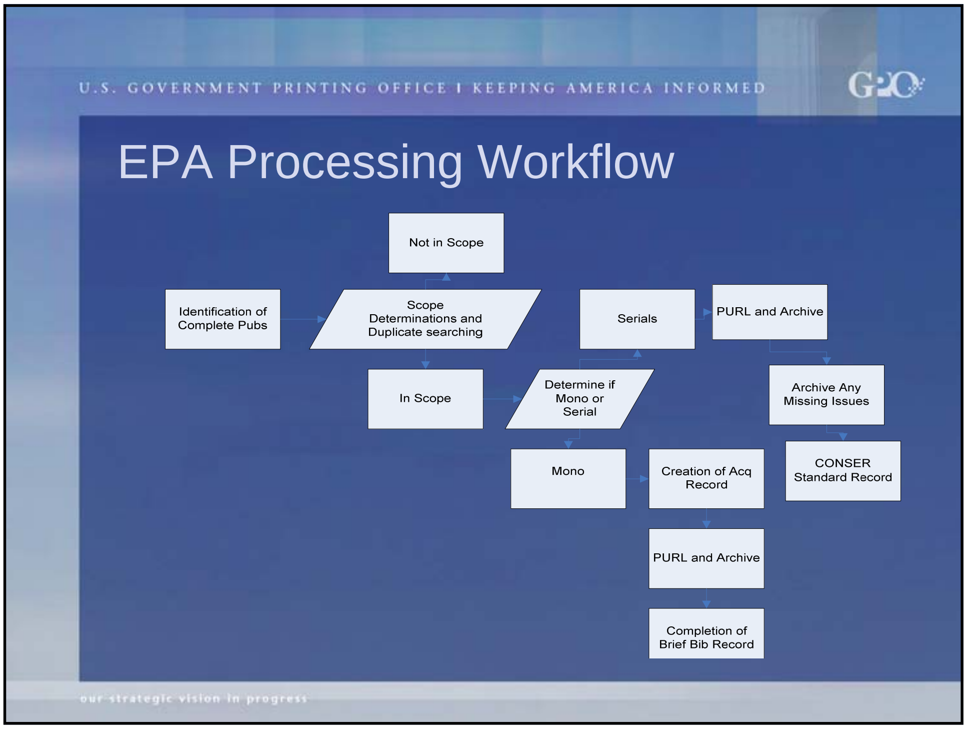
Sample of 300 Results

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



Processing Times for the Sample

Processing Step	Average Time
Identification of a Complete Publication	2 minutes
Scope Determination and Search for Duplicates	17 minutes
Creation of Brief Bibliographic Record	30 minutes
Creation of CONSER Standard Record	2 hours 30 minutes
Add PURL to Publications Distributed in Tangible Format	7.5 minutes
Creation of Browse Table	4 hours total to create the entire browse table



EPA Processing Workflow

Identification of Complete Pubs

Scope Determinations and Duplicate searching

Not in Scope

In Scope

Determine if Mono or Serial

Serials

PURL and Archive

Archive Any Missing Issues

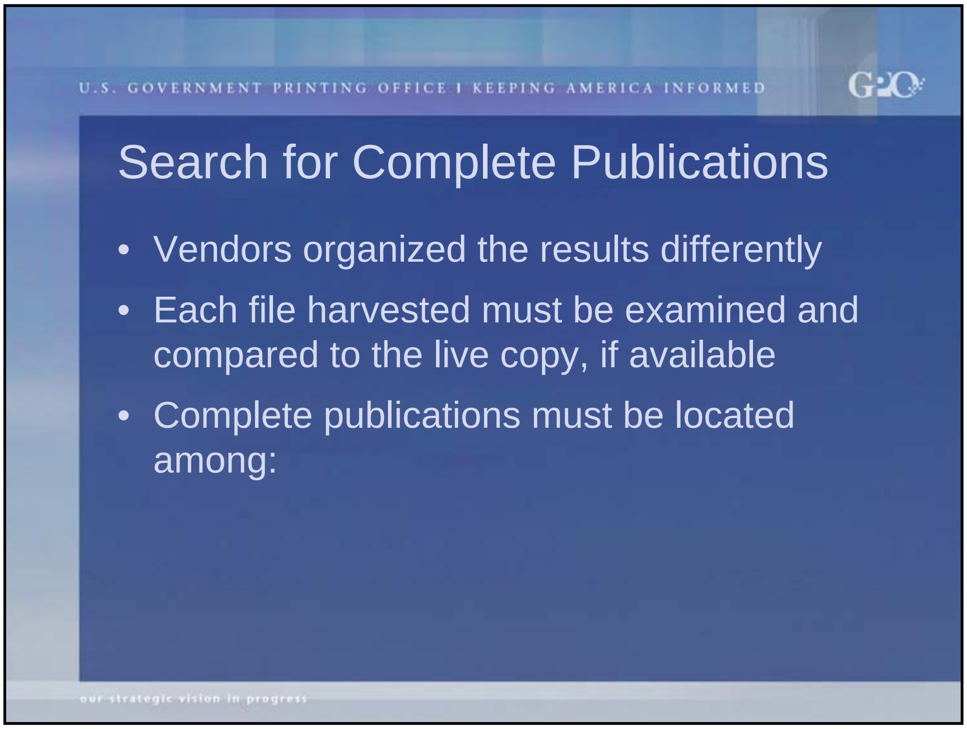
CONSER Standard Record

Mono

Creation of Acq Record

PURL and Archive

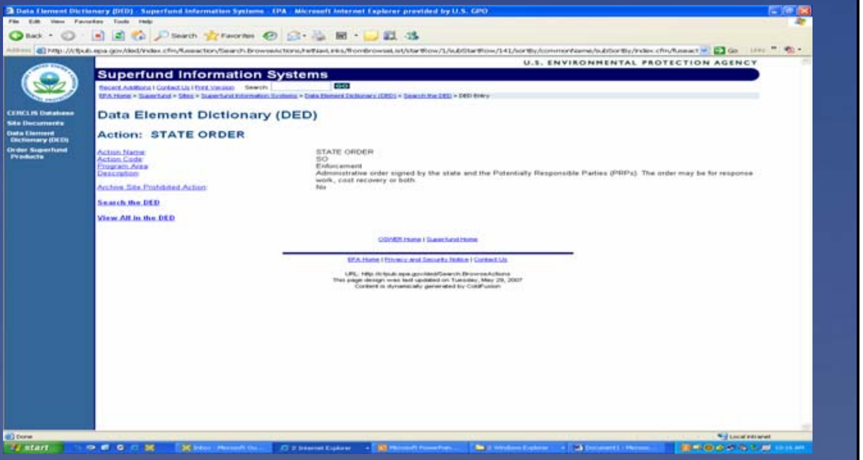
Completion of Brief Bib Record



Search for Complete Publications

- Vendors organized the results differently
- Each file harvested must be examined and compared to the live copy, if available
- Complete publications must be located among:

Database Results


Data Element Dictionary (DED) - Superfund Information Systems - EPA - Microsoft Internet Explorer provided by U.S. GPO
File Edit View Favorites Tools Help
Back
Search
Favorites
U.S. ENVIRONMENTAL PROTECTION AGENCY
Superfund Information Systems
Recent Additions | Contact Us | Print Version
Search:
EPA Home > Superfund > Sites > Superfund Information Systems > Data Element Dictionary (DED) > Search the DED > DED Entry
CERCLIS Database
Site Documents
Data Element Dictionary (DED)
Order Superfund Products

Data Element Dictionary (DED)

Action: STATE ORDER

Action Name:	STATE ORDER
Action Code:	SO
Program Area:	Enforcement
Description:	Administrative order signed by the state and the Potentially Responsible Parties (PRPs). The order may be for response work, cost recovery or both.
Archive Site Prohibited Action:	No

Search the DED

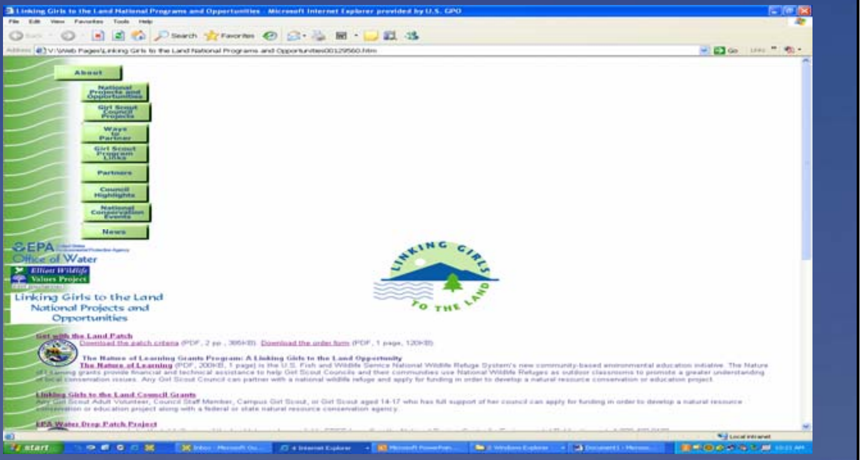
View All in the DED

OSWER Home | Superfund Home

EPA Home | Privacy and Security Notice | Contact Us

URL: http://cfpub.epa.gov/ded/Search.BrowseActions
This page design was last updated on Tuesday, May 29, 2007
Content is dynamically generated by ColdFusion

Web Pages


Linking Girls to the Land National Programs and Opportunities - Microsoft Internet Explorer provided by U.S. GPO
File Edit View Favorites Tools Help
Search
Favorites
V:\Web Pages\Linking Girls to the Land National Programs and Opportunities00129560.htm
About
National Projects and Opportunities
Girl Scout Council Projects
Ways to Partner
Girl Scout Program Links
Partners
Council Highlights
National Conservation Events
News
EPA
Office of Water
Elliott Wildlife Values Project
LINKING GIRLS TO THE LAND
Local intranet
start

Linking Girls to the Land National Projects and Opportunities

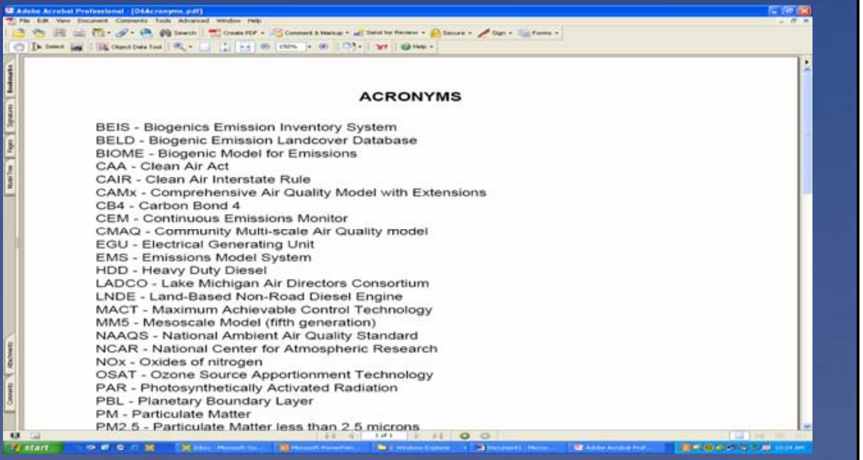
Get with the Land Patch
Download the patch criteria (PDF, 2 pp., 395KB) Download the order form (PDF, 1 page, 120KB)

The Nature of Learning Grants Program: A Linking Girls to the Land Opportunity
The Nature of Learning (PDF, 200KB, 1 page) is the U.S. Fish and Wildlife Service National Wildlife Refuge System's new community-based environmental education initiative. The Nature of Learning grants provide financial and technical assistance to help Girl Scout Councils and their communities use National Wildlife Refuges as outdoor classrooms to promote a greater understanding of local conservation issues. Any Girl Scout Council can partner with a national wildlife refuge and apply for funding in order to develop a natural resource conservation or education project.

Linking Girls to the Land Council Grants
Any Girl Scout Adult Volunteer, Council Staff Member, Campus Girl Scout, or Girl Scout aged 14-17 who has full support of her council can apply for funding in order to develop a natural resource conservation or education project along with a federal or state natural resource conservation agency.

EPA Water Drop Patch Project

Partially Harvested Publications

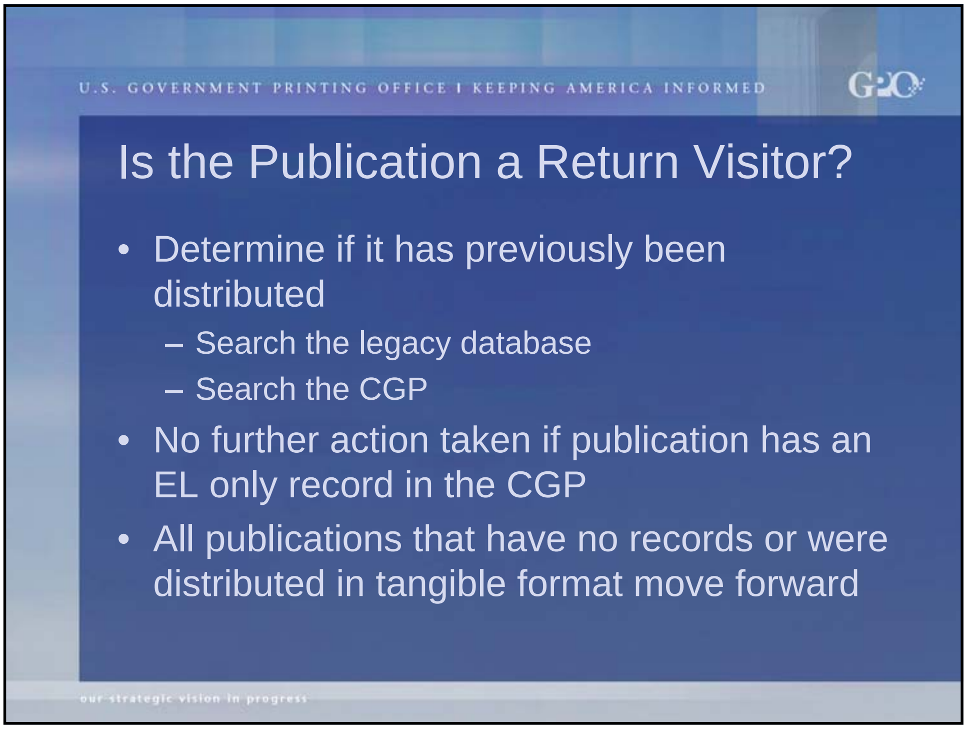

Adobe Acrobat Professional - [06Acronyms.pdf]
File Edit View Document Comments Tools Advanced Window Help
Search
Create PDF
Comment & Markup
Send for Review
Secure
Sign
Forms
Select
Object Data Tool
Help
start

ACRONYMS

BEIS - Biogenics Emission Inventory System
BELD - Biogenic Emission Landcover Database
BIOME - Biogenic Model for Emissions
CAA - Clean Air Act
CAIR - Clean Air Interstate Rule
CAMx - Comprehensive Air Quality Model with Extensions
CB4 - Carbon Bond 4
CEM - Continuous Emissions Monitor
CMAQ - Community Multi-scale Air Quality model
EGU - Electrical Generating Unit
EMS - Emissions Model System
HDD - Heavy Duty Diesel
LADCO - Lake Michigan Air Directors Consortium
LNDE - Land-Based Non-Road Diesel Engine
MACT - Maximum Achievable Control Technology
MM5 - Mesoscale Model (fifth generation)
NAAQS - National Ambient Air Quality Standard
NCAR - National Center for Atmospheric Research
NOx - Oxides of nitrogen
OSAT - Ozone Source Apportionment Technology
PAR - Photosynthetically Activated Radiation
PBL - Planetary Boundary Layer
PM - Particulate Matter
PM2.5 - Particulate Matter less than 2.5 microns

Does the Publication Belong?

- Determine if publication is within scope of the FDLP and/or the C&I Program
- Some of the questions considered:
  - Is it published by a US Government agency?
  - Is the information in it covered by copyright?
  - Who is the copyright holder?
  - Is the primary source of funding for it government money?
  - Does it contain data that may violate a citizen's privacy?

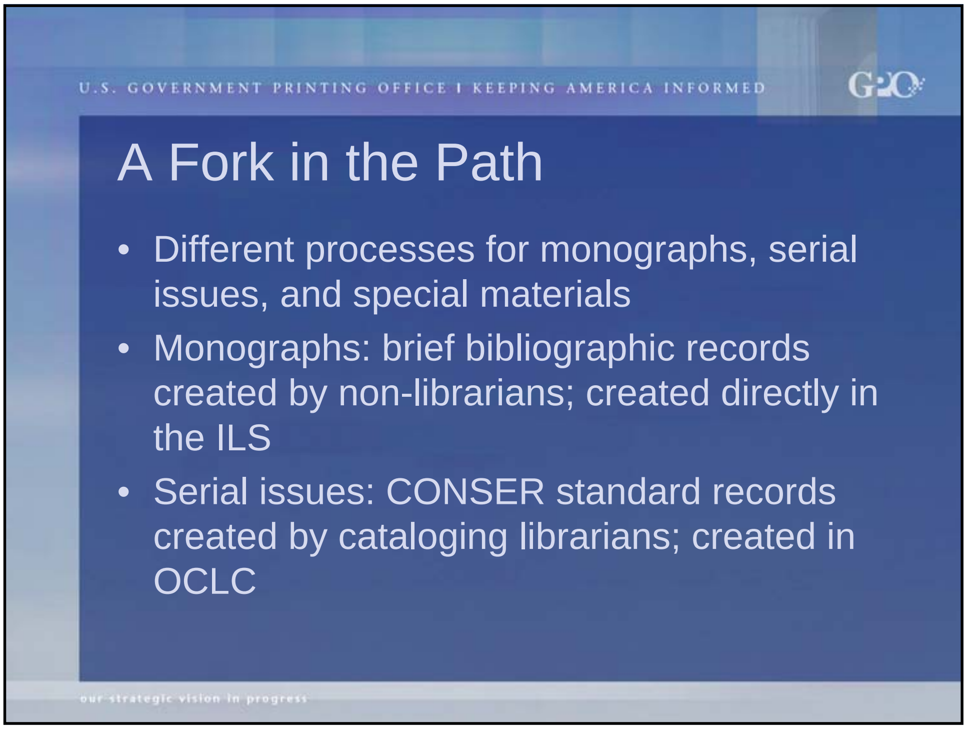


U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

Is the Publication a Return Visitor?

- Determine if it has previously been distributed
  - Search the legacy database
  - Search the CGP
- No further action taken if publication has an EL only record in the CGP
- All publications that have no records or were distributed in tangible format move forward

our strategic vision in progress



U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

A Fork in the Path

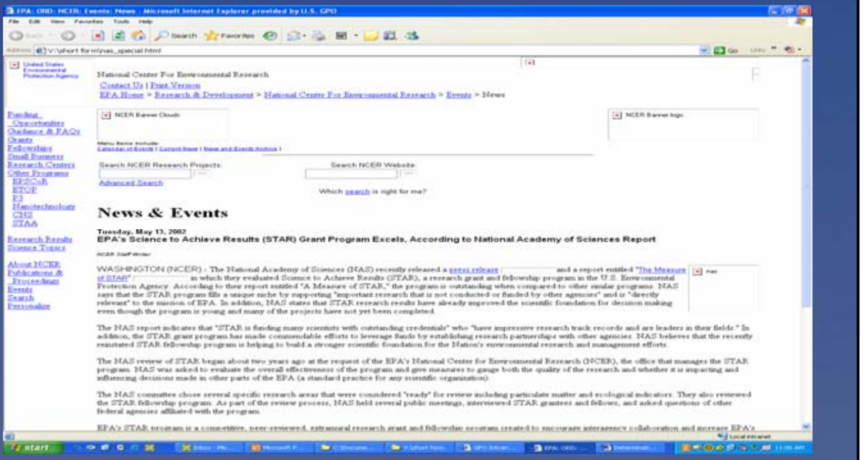
- Different processes for monographs, serial issues, and special materials
- Monographs: brief bibliographic records created by non-librarians; created directly in the ILS
- Serial issues: CONSER standard records created by cataloging librarians; created in OCLC

our strategic vision in progress

Special Materials

- Special materials include:
  - News releases
  - Transmittals
  - Forms
  - Announcements
- Included in List of Special Materials in the Monthly Catalog
- Investigating methods to provide bibliographic access to this material

Special Material


EPA: ORD: NCER: Events: News - Microsoft Internet Explorer provided by U.S. GPO
File Edit View Favorites Tools Help
Search Favorites
Address V:\short form\nas_special.html
Go
United States Environmental Protection Agency
NCER Banner Clouds
NCER Banner logo
start

National Center For Environmental Research

Contact Us | Print Version

EPA Home > Research & Development > National Center For Environmental Research > Events > News

Funding Opportunities
Guidance & FAQs
Grants
Fellowships
Small Business
Research Centers
Other Programs
EPSCoR
ETOP
P3
Nanotechnology
CHS
STAA
Research Results
Science Topics
About NCER
Publications & Proceedings
Events
Search
Personalize

Menu items include:
Calendar of Events | Current News | News and Events Archive |

Search NCER Research Projects:

Search NCER Website:

Advanced Search

Which search is right for me?

News & Events

Tuesday, May 13, 2002

EPA's Science to Achieve Results (STAR) Grant Program Excels, According to National Academy of Sciences Report

NCER Staff Writer

WASHINGTON (NCER) - The National Academy of Sciences (NAS) recently released a press release and a report entitled "The Measure of STAR" in which they evaluated Science to Achieve Results (STAR), a research grant and fellowship program in the U.S. Environmental Protection Agency. According to their report entitled "A Measure of STAR," the program is outstanding when compared to other similar programs. NAS says that the STAR program fills a unique niche by supporting "important research that is not conducted or funded by other agencies" and is "directly relevant" to the mission of EPA. In addition, NAS states that STAR research results have already improved the scientific foundation for decision making even though the program is young and many of the projects have not yet been completed.

The NAS report indicates that "STAR is funding many scientists with outstanding credentials" who "have impressive research track records and are leaders in their fields." In addition, the STAR grant program has made commendable efforts to leverage funds by establishing research partnerships with other agencies. NAS believes that the recently reinstated STAR fellowship program is helping to build a stronger scientific foundation for the Nation's environmental research and management efforts.

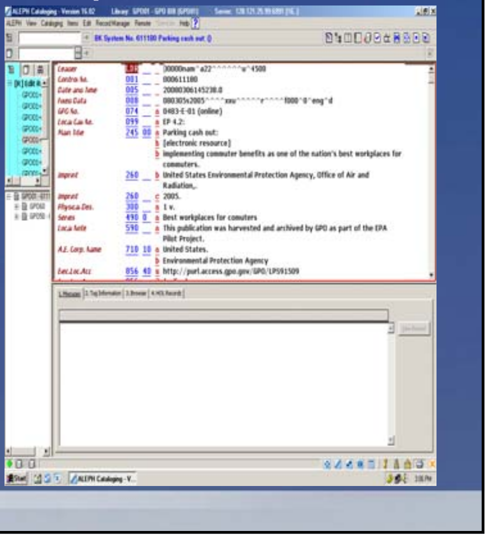
The NAS review of STAR began about two years ago at the request of the EPA's National Center for Environmental Research (NCER), the office that manages the STAR program. NAS was asked to evaluate the overall effectiveness of the program and give measures to gauge both the quality of the research and whether it is impacting and influencing decisions made in other parts of the EPA (a standard practice for any scientific organization).

The NAS committee chose several specific research areas that were considered "ready" for review including particulate matter and ecological indicators. They also reviewed the STAR fellowship program. As part of the review process, NAS held several public meetings, interviewed STAR grantees and fellows, and asked questions of other federal agencies affiliated with the program.

EPA's STAR program is a competitive, peer-reviewed, extramural research grant and fellowship program created to encourage interagency collaboration and increase EPA's

Monograph: 1st Step

- Begin brief bibliographic record in ILS
- Includes basic bibliographic information and the SuDoc stem
- Not accessible in CGP
- Created by staff in Content Acquisitions



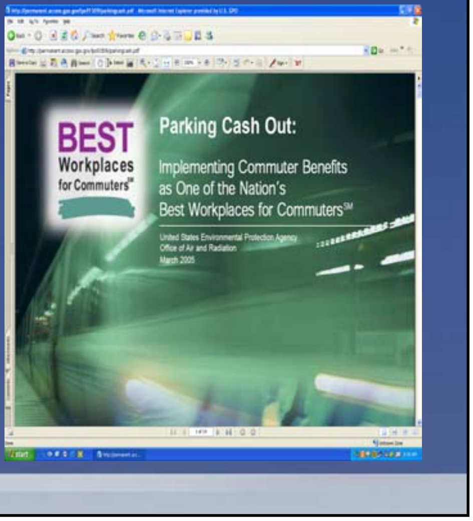
BK System No. 611100 Parking cash out ()

```
Leader          LDR        00000nam^a22^^^^^^^a^4500
Control No.     001        000611100
Date and Time   005        20080306145238.0
Fixed Data      008        080305s2005^^^^xxu^^^^^^r^^^^f000^0^eng^d
GPO No.         074      a 0483-E-01 (online)
Local Call No.  099      a EP 4.2:
Main Title      245 00   a Parking cash out:
                         b [electronic resource]
                         b implementing commuter benefits as one of the nation's best workplaces for
                           commuters.
Imprint         260      b United States Environmental Protection Agency, Office of Air and
                           Radiation,.
Imprint         260      c 2005.
Physical Des.   300      a 1 v.
Series          490 0    a Best workplaces for comuters
Local Note      590      a This publication was harvested and archived by GPO as part of the EPA
                           Pilot Project.
A.E. Corp. Name 710 10   a United States.
                         b Environmental Protection Agency
Elec.Loc.Acc.   856 40   u http://purl.access.gpo.gov/GPO/LPS91509
```

ALEPH Cataloging - Version 16.02
ALEPH View Cataloging Items Edit RecordManager Remote Help
1.Message 2.Tag Information 3.Browse 4.HOL Records
ALEPH Cataloging - V...

Monograph: 2nd Step

- Archive publication on the Permanent Server
- Create PURL to issuing agency's site or the archived copy
- Add PURL to bibliographic record



BEST Workplaces for Commuters℠

Parking Cash Out:

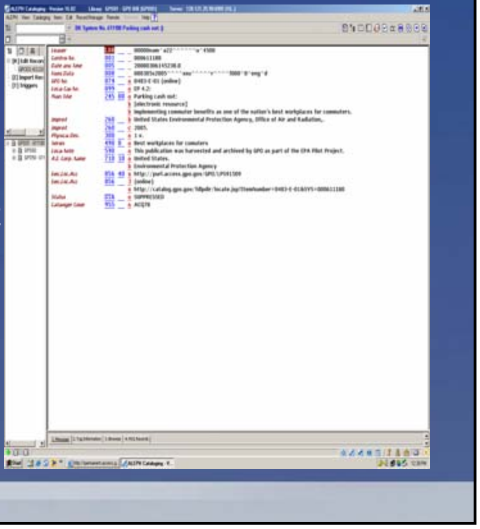
Implementing Commuter Benefits as One of the Nation's Best Workplaces for Commuters℠

United States Environmental Protection Agency
Office of Air and Radiation
March 2005

U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

Monograph: Final Step

- Complete brief bibliographic record
- Validate SuDoc and item numbers
- Record indicates is part of EPA Pilot Project
- Now accessible in CGP
- Finalized by staff in Library Technical Information Support



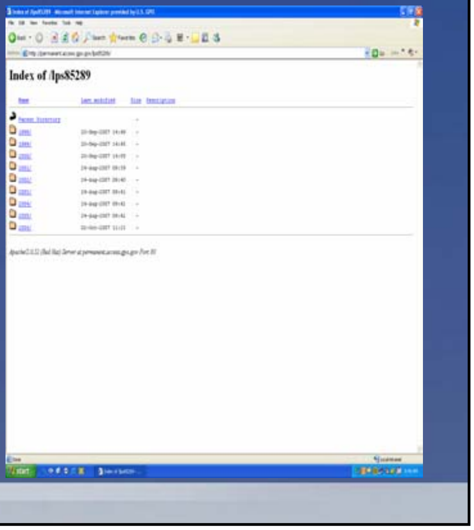
```
Leader          000       00000nam a22       a 4500
Control No.     001       000611180
Date and Time   005       20080306145238.0
Fixed Data      008       080305s2005    xxu          s       000 0 eng d
GPO No.         074    a  0483-E-01 (online)
Local Call No.  099    a  EP 4.2:
Main Title      245 00 a  Parking cash out:
                       h  [electronic resource]
                       b  implementing commuter benefits as one of the nation's best workplaces for commuters.
Imprint         260    b  United States Environmental Protection Agency, Office of Air and Radiation,.
Imprint         260    c  2005.
Physical Des.   300    a  1 v.
Series          490 0  a  Best workplaces for commuters
Local Note      590    a  This publication was harvested and archived by GPO as part of the EPA Pilot Project.
Add. Corp. Name 710 10 a  United States.
                       b  Environmental Protection Agency
Elec.Loc.Acc    856 40 u  http://purl.access.gpo.gov/GPO/LPS91509
Elec.Loc.Acc    856    3  (online)
                       u  http://catalog.gpo.gov/fdlpdir/locate.jsp?ItemNumber=0483-E-01&SYS=000611180
Status          916    a  SUPPRESSED
Cataloger Code  955    a  ACQ78
```

our strategic vision in progress

U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

Serial Issue: 1st Step

- Archive issue on the Permanent Server
- Harvest and archive additional issues
- Create PURL to issuing agency's site or the archived copy



http://permanent.access.gpo.gov/lps85289/

Index of /lps85289

Name	Last modified	Size	Description
Parent Directory		-	
1998/	20-Sep-2007 14:49	-	
1999/	20-Sep-2007 14:48	-	
2000/	20-Sep-2007 14:55	-	
2001/	29-Aug-2007 09:39	-	
2002/	29-Aug-2007 09:40	-	
2003/	29-Aug-2007 09:41	-	
2004/	29-Aug-2007 09:42	-	
2005/	29-Aug-2007 09:42	-	
2006/	30-Oct-2007 11:22	-	

Apache/2.0.52 (Red Hat) Server at permanent.access.gpo.gov Port 80

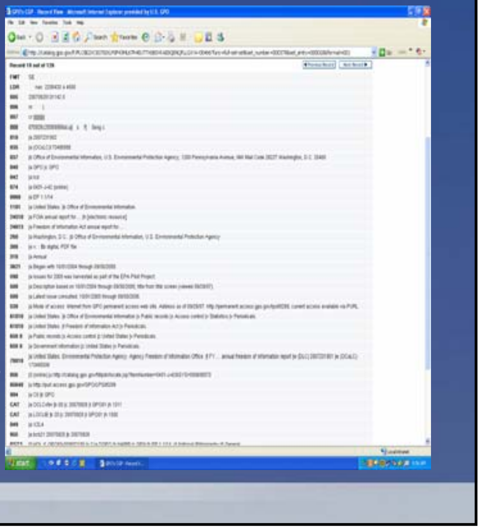
our strategic vision in progress

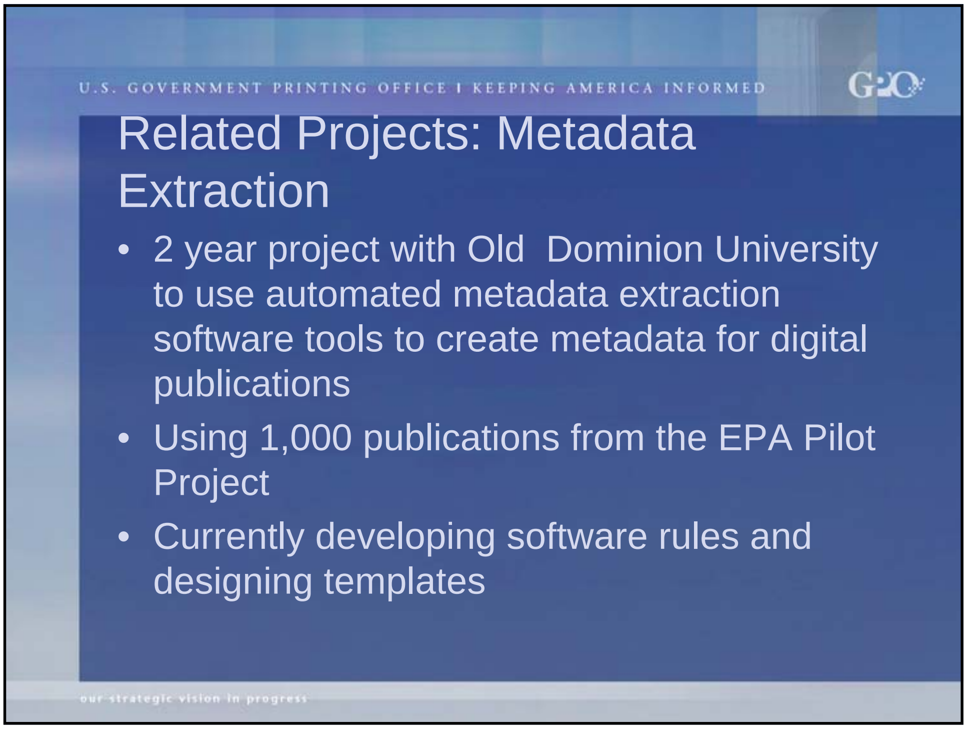
Serial Issue: 2nd Step

- Issuing agency contacted for information, if needed
  - Is the serial still being published
  - Frequency
- Can take time to get a response
- New SuDoc class and item number created, if needed

Serial Issue: Final Step

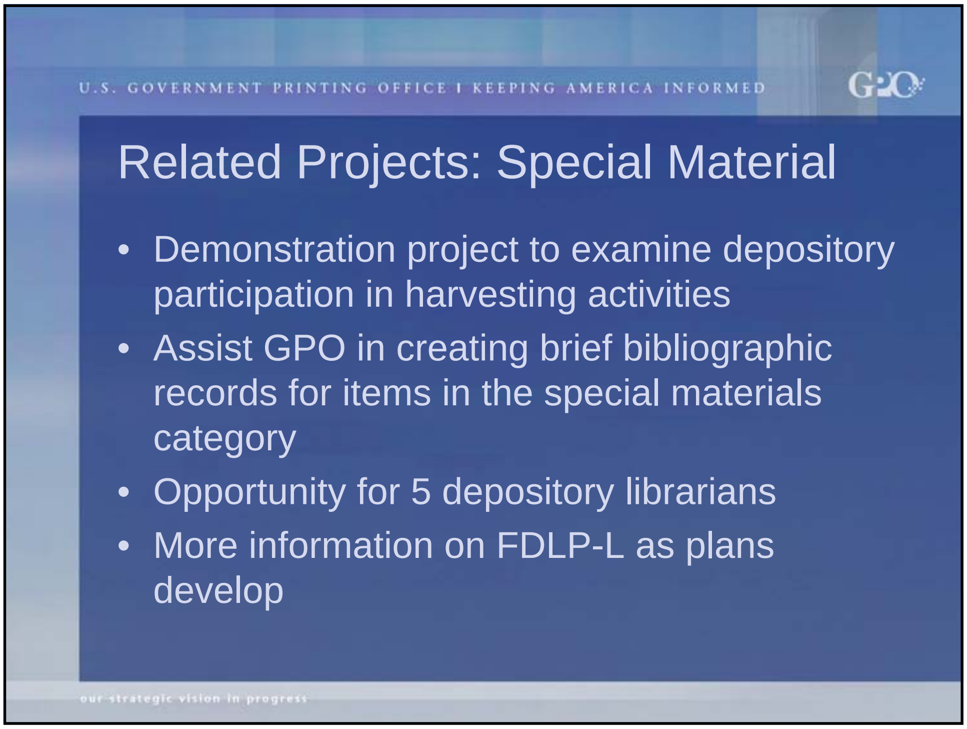
- Create CONSER standard record in OCLC
- Load record into the ILS
- Record in ILS indicates is part of the EPA Pilot Project
- Now accessible in CGP





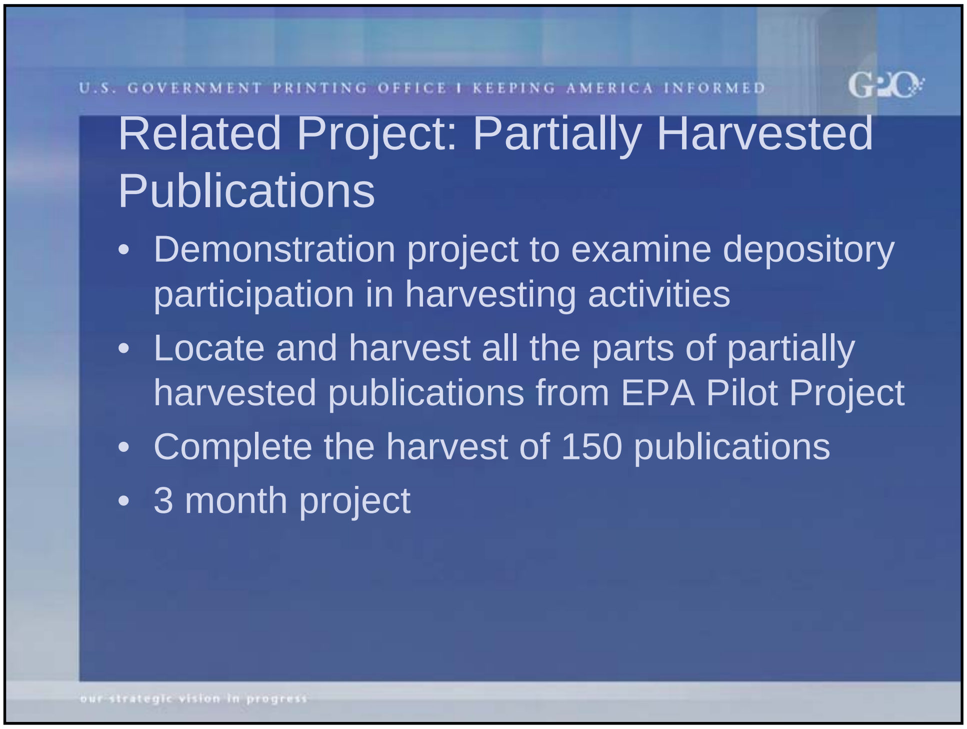
Related Projects: Metadata Extraction

- 2 year project with Old Dominion University to use automated metadata extraction software tools to create metadata for digital publications
- Using 1,000 publications from the EPA Pilot Project
- Currently developing software rules and designing templates



Related Projects: Special Material

- Demonstration project to examine depository participation in harvesting activities
- Assist GPO in creating brief bibliographic records for items in the special materials category
- Opportunity for 5 depository librarians
- More information on FDLP-L as plans develop

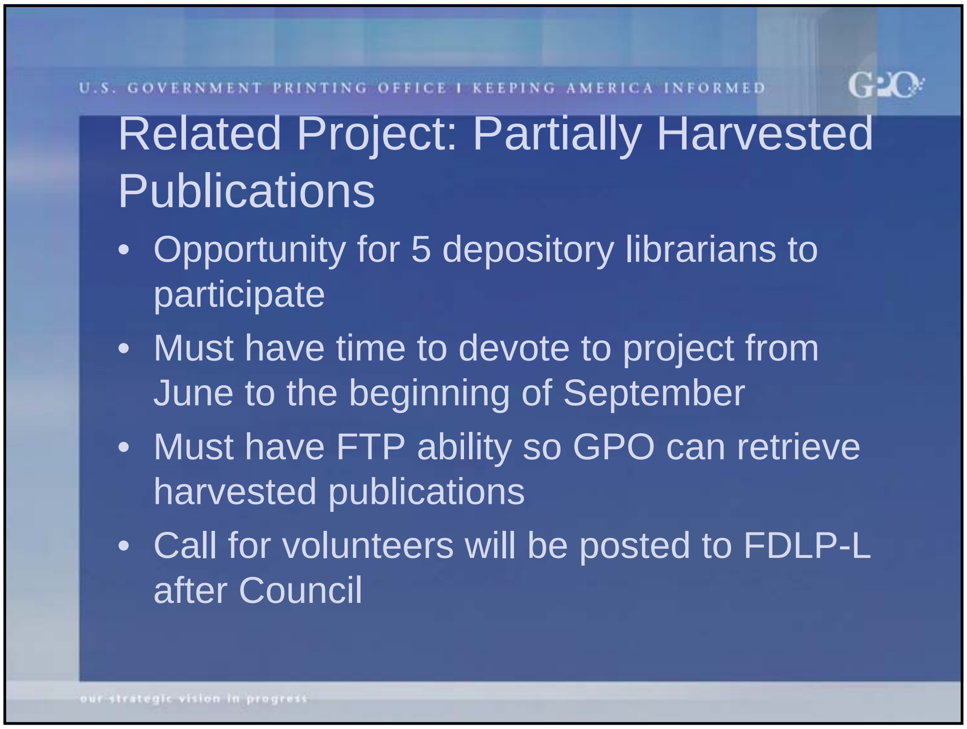


U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

Related Project: Partially Harvested Publications

- Demonstration project to examine depository participation in harvesting activities
- Locate and harvest all the parts of partially harvested publications from EPA Pilot Project
- Complete the harvest of 150 publications
- 3 month project

our strategic vision in progress

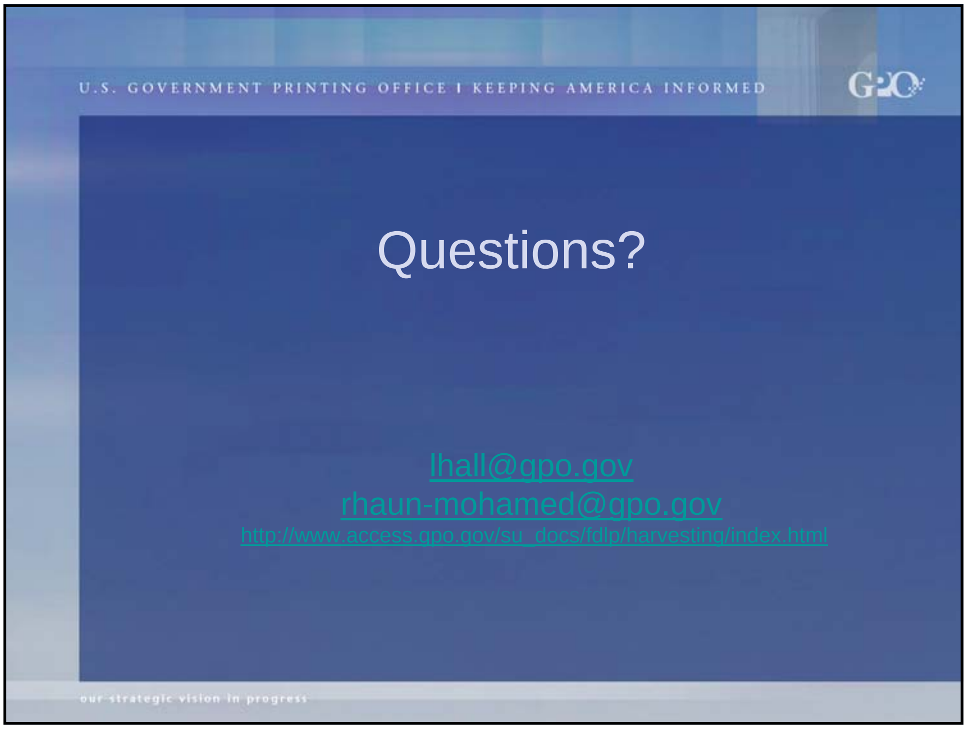


U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

Related Project: Partially Harvested Publications

- Opportunity for 5 depository librarians to participate
- Must have time to devote to project from June to the beginning of September
- Must have FTP ability so GPO can retrieve harvested publications
- Call for volunteers will be posted to FDLP-L after Council

our strategic vision in progress



U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

Questions?

lhall@gpo.gov

rhaun-mohamed@gpo.gov

http://www.access.gpo.gov/su_docs/fdlp/harvesting/index.html

our strategic vision in progress

Web Harvesting

Web Harvesting Update for the Depository Library Council

BACKGROUND

GPO has been harvesting digital Federal publications since the late 1990's. These publications are acquired using two harvesting techniques: manual and semi-manual. Manual harvesting is the capture of known digital publications through manual identification and the saving of all associated publication files, while semi-manual uses a software tool to schedule the content capture and re-harvesting of known content at known Web sites.

In 2006, GPO, with the approval of the U.S. Environmental Protection Agency (EPA), undertook an automated pilot project to harvest materials from EPA Web sites. Using two different vendors, over 200,000 files were acquired. Subsequent action and review of this material has focused on ways to identify efficient methods to bring the files that are within scope of the FDLP and the Cataloging and Indexing (C&I) Program under bibliographic control.

SAMPLE PUBLICATIONS FROM THE RESULTS OF THE EPA PILOT PROJECT

Library Services and Content Management (LSCM) staff processed a sample of 300 publications harvested during the EPA Pilot Project. The purpose of working through this sample was to determine workflow and staffing implications, as well as to estimate the amount of time that would be required to process all of the publications acquired during the EPA Pilot Project.

Two mechanisms were used for making accessible the publications found to be within scope of the FDLP. The depository community was asked to provide feedback on both. The majority of publications were made accessible through bibliographic records in the Catalog of U.S. Government Publications (CGP). Monographs were cataloged using the new brief bibliographic record format, while serials were cataloged following the CONSER standard record format. At the request of the Depository Library Council (DLC), LSCM also explored a mechanism that enables public access to Web harvested content while these publications are in the queue for brief bibliographic records, and posted a small portion of the sample to *GPO Access* using a browse table. Publications made accessible through this mechanism will be cataloged in the CGP in the future.

Of the 300 publications identified for the sample, 62% had not previously been distributed through the FDLP. Based on the average processing times, it will take 49 minutes to process a single monograph from identification in the vendors' results to the finalization of the brief bibliographic record. Average processing time for a serial title is two hours and 49 minutes, from identification through the creation of the CONSER standard record. The complete analysis of the sample is available at **<http://www.access.gpo.gov/su_docs/fdlp/harvesting/index.html>**.

ONGOING HARVESTING ACTIVITIES

GPO continues to take an active role in Web harvesting efforts. As a member of CENDI's Web Harvesting Task Group, GPO is participating in discussions with other Federal agencies that will lead to the development of a white paper on Web harvesting best practices. Other agencies participating in the Task Group include NASA, the EPA, NTIS, and the Library of Congress. CENDI is an interagency working group of senior scientific and technical information managers from thirteen Federal agencies. Any white paper or other products resulting from this interagency effort will be shared with DLC and the depository library community. This information will also be used to further develop requirements as appropriate for the Web harvesting components of the FDsys.

Manual Web harvesting efforts also continue on a daily basis. Staff in Library Technical Information Services (LTIS) are responsible for identifying, classifying, and archiving new digital publications as part of their daily duties. These publications are included in the monthly New Electronic Titles report. While the bulk of the harvesting is performed manually, GPO staff continue to work with a semi-automated harvester to acquire issues of serials and publications in non-PDF formats.

In addition, GPO staff have begun another project to provide more access to the files acquired during the EPA Pilot Project. Following a workflow established based on the results of the sample of 300 publications, staff will process 500 files in the coming months. Many of these files will either be incomplete or not within scope of the FDLP or the C&I Program, so the depository community should not expect to locate 500 additional bibliographic records in the CGP after the completion of this project.

HARVESTING ISSUES

- Most of the material acquired during the EPA Pilot Project is still available from EPA Web sites and has not been lost. Some of the material retrieved was not in-scope, not a complete publication, or both. GPO is developing a demonstration project with libraries to assist in identifying where the complete publications can be located.
- The sheer number of files identified and retrieved during the EPA Pilot Project is overwhelming. In addition to the development of automated metadata extraction tools to assist with bibliographic control over this material, GPO is developing a demonstration project for libraries to assist in the creation of brief cataloging records for a subset of the results from the EPA Pilot Project.
- Many of the files returned from the EPA crawls were parts of a database. GPO continues to focus Web harvesting efforts on Federal Government information products or publications as identified under Title 44. Other agencies may have a different role and approach, including the downloading of entire or partial Web sites.

DEPOSITORY PARTICIPATION IN HARVESTING EFFORTS

The Federal depository library community has expressed interest in participating in LSCM's efforts to process these publications. In order to determine how best to utilize the depository libraries' offer of assistance, LSCM will undertake a demonstration project with ten depositories. The volunteer depositories will be asked to assist LSCM either in locating the missing sections of partially harvested publications or by creating brief bibliographic records for publications in the Special Materials category, which includes press releases, transmittals, notices, and forms. Additional information about these demonstration projects will be shared via FDLP-L and the FDLP Desktop.

RELATED PROJECTS

One thousand monographs within scope of the FDLP have been identified from the EPA Pilot Project for inclusion in the Automated Metadata Extraction Project. This two-year project with the Defense Technical Information Service (DTIC) and Old Dominion University (ODU) will use automated metadata extraction software tools to create metadata for groups of electronic publications in GPO's electronic collection. GPO expects to receive the results near the end of the project.

ASSUMPTIONS

- GPO will continue to participate in Web harvesting efforts to obtain in-scope material for the FDLP and the C&I Program as required under 44 USC.
- GPO is bound by the Congressional appropriations for the Salaries & Expenses funding requirements for the FDLP and C&I Programs.
- All materials identified for inclusion in the FDLP must be brought under bibliographic control as directed under the C&I Program.
- GPO does not have gift authority to either give funding or gifts or to receive them. All partnerships must represent a contribution of an equal exchange between all parties.
- Automated Web harvesting initiatives will become systematic as part of release 2 of the FDsys.
- Materials harvested under the EPA Pilot Project are being made available as staff time and processing permit. Completion of the processing of this material will necessarily require an automated metadata extraction process that does not yet exist.