U.S. GOVERNMENT
PRINTING OFFICE
KEEPING AMERICA INFORMED

CRITERIA AND PARAMETERS FOR
GPO’S WEB HARVESTING PILOT PROJECT

TABLE OF CONTENTS

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	