


The FDLP and Web Harvesting

Permanent Access to Online Federal Resources



GPO U.S. GOVERNMENT PRINTING OFFICE
Keeping America Informed | www.gpo.gov





Web Harvesting Defined

- [Web scraping](#) from many sites
- Implementation of a [Web crawler](#) that uses human expertise or machine guidance to direct itself to [URLs](#) composing a specialized collection or set of knowledge
- For GPO, Web harvesting can be thought of as focused or directed Web crawling that is non-invasive





FEDERAL DEPOSITORY LIBRARIES
Free Information, Dedicated Service, Limitless Possibilities

What GPO is Doing Now

- Harvesting piece level titles that fit within the scope of the FDLP
- Using semi-automated and manual harvesting tools to capture born digital content
- Archiving the harvested content using redundant storage
- Providing access to Web harvested material through the CGP and search engines that are indexing the servers
- Assigning PURLs for almost all Web harvested material



FEDERAL DEPOSITORY LIBRARIES
Free Information, Dedicated Service, Limitless Possibilities

Harvesting Challenges

- Publications vs. Web pages
- PURLing resources: exceptions: Databases; publications on FDsys and GPO Access; multimedia formats; large or resource intensive files (National Map example)
- Metadata needs: bibliographic requirements and technical requirements
- Applications within publications





FEDERAL DEPOSITORY LIBRARIES
Free Information, Dedicated Service, Limitless Possibilities

GPO's Path Forward



- Continue investigation and review of Web harvesting best practices
- Identification of test sets for automated harvest and ingest into FDSys
- Increased partnership activity with agencies for access to born digital publications



FEDERAL DEPOSITORY LIBRARIES
Free Information, Dedicated Service, Limitless Possibilities

Questions?