U.S. GOVERNMENT PRINTING OFFICE I KEEPING AMERICA INFORMED

GPO Harvesting Pilot

Kirk Knoll

April 3, 2006

Harvesting Pilot Background

- Statutory Obligation (FDLP)
- Digital Age
- Fugitive Documents

- The PMO is working closely with Library Services and Collection Management to manage the pilot.

Two concurrent GPO contracts with…

- IIA
- Blue Angel Technologies

Pilot agency: EPA

The Pilot Schedule

- 3 separate crawls beginning April, 2006
- Ending October, 2006

Key Deliverables

- Rules and instructions
- Comparison of harvested collection with the existing FDLP collection
- Harvested EPA content within scope of the FDLP

Harvesting Challenges

- Writing rules
  - In scope
  - Boundaries
    - How far does the harvester go down the rabbit hole?
    - Where does a publication start and stop?

The Harvester will be comprised of 3 tools

- Discovery tools: to locate content on a federal agency web site
- Assessment tools: to determine whether discovered content is within scope of FDLP using rules and instructions
- Harvesting tools: to capture and gather in-scope content.

KEY Take-Aways

- Information necessary to further refine GPO's requirements for a comprehensive harvesting solution to be implemented into FDsys.
- Information will be shared with the FDsys Master Integrator (MI)
- GPO and the MI will collaboratively select the best technology to perform the FDsys harvesting functions.