U.S. GOVERNMENT PRINTING OFFICE I KEEPING AMERICA INFORMED

# GPO Harvesting Pilot

## Federal Depository Library Council Fall Meeting
## October 23, 2006



# Harvesting Project Background

- Statutory Obligation and Mission
- Digital Age
- Goal of Comprehensive Collection



# The solution?
## Automated web harvesting

The goal of the harvesting tool is to retrieve publications that are within scope of GPO dissemination programs.

# Future Digital System (FDsys) Submission Reference Model

## Content Submission

- Converted Content
- **Harvested Content**
- Deposited Content
- Style Tools

# The Harvester will be comprised of 3 tools

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

# Pilot Project Overview

- Recognized automated harvesting as a need.
- Aligned project goals and objectives to FDsys.
- Issued Request for Proposals (RFP) in late 2005, made award in early 2006.

# Approach

- Concurrent contracts with vendors experienced with harvesting:
  - Information International Associates
  - Blue Angel Technologies

## Pilot agency: EPA

# Schedule

- Pilots began in April, 2006 and contractor work was completed in early October, 2006.
- Three separate crawls of the EPA Web site.

# Key Deliverables

- Rules used to determine scope
- Comparison of harvested collection with existing cataloging records
- Harvested EPA content within scope

# Pilot Process

- GPO provided both contractors with initial set of criteria and parameters for determining scope.
- Contractors wrote rules based on information provided by GPO.
- Crawled and harvested metadata as well as content.
- After each crawl, GPO analyzed and reviewed results and worked with contractors to modify rules.
- Selected EPA databases identified by GPO were harvested for the second and third crawls.

# Contractor Methodologies: IIA

- Had previous knowledge of the EPA Web site.
- A major focus was categorization of content and development of rules associated with each category.
- Created data mining and other tools that analyzed content for patterns of key terms and content characteristics.

# Contractor Methodologies: Blue Angel

- Limited knowledge of the EPA Web site.
- Focused on writing rules that:
  - Exclude types of documents that are known not to be in scope.
  - Include documents that contain keywords in specific sections of documents (e.g., metadata, front matter, title, header and footer)

# Harvester Rules (Examples)

- Rules that determine whether a discovered document is an in-scope publication:
  - Exclude: internal memos, work in progress drafts, fragments, purchase orders, Statements of Work.
  - Include: documents that contain specific words or phrases (e.g., chapter, appendix, technical report, etc.) that indicate content is a publication.
  - Link analysis.
  - Indicate authorship or funding by an official Government source.

# Harvester Rules (Examples)

- Rules that determine whether a publication is an EPA publication:
  - Excluded any documents that did not include reference to EPA.
  - Excluded documents not published by EPA (e.g., sections from Federal Register, CFR, etc.)
  - Included documents that contained indicators that they are EPA publications (e.g., title, author, description)

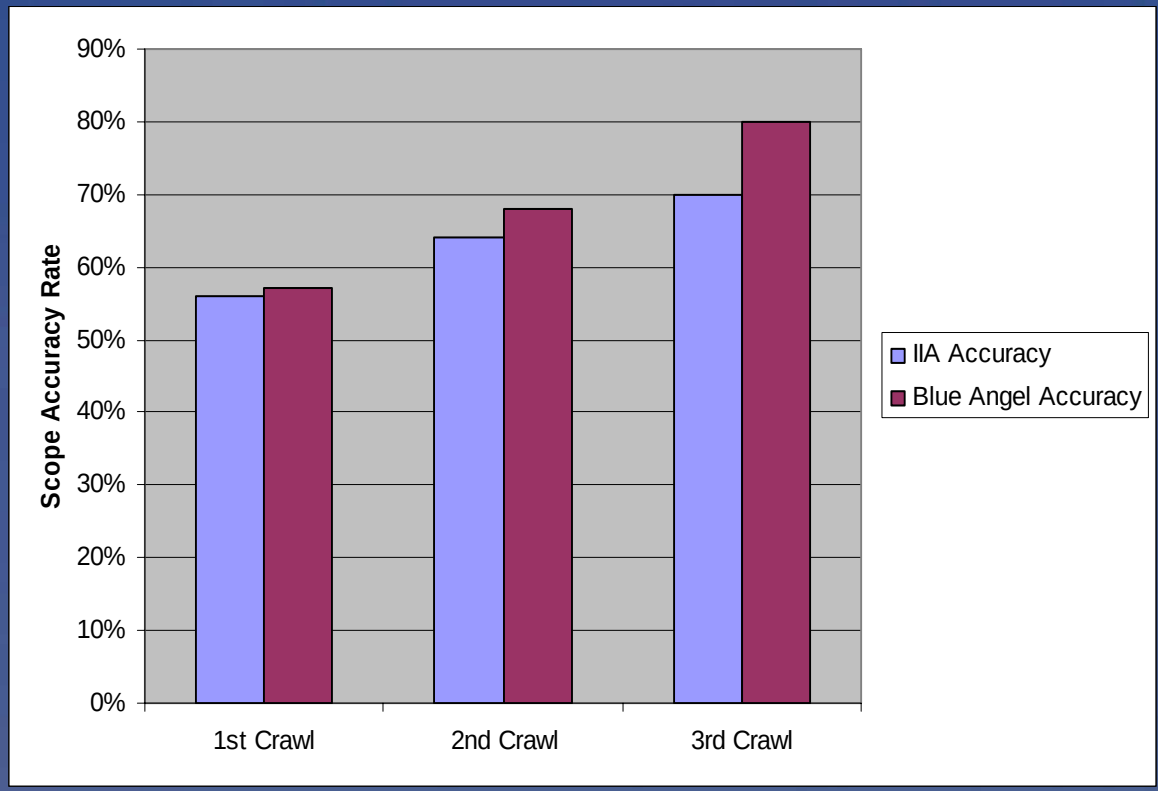
# Blue Angel Technologies Pilot Results

- Discovered and harvested 83,229 documents that their technologies and rules have deemed to be in scope.
  - "Documents" refers to complete publications or individual parts of publications.
- The accuracy rate of scope determination is projected to be between 75% and 85% based on initial GPO sampling and analysis.

# Information International Associates Pilot Results

- Discovered and harvested 239,478 documents that their technologies and rules have deemed to be in scope.
  - "Documents" refers to complete publications or individual parts of publications.
- The accuracy rate of scope determination is projected to be between 70% and 75% based on initial GPO sampling and analysis.

# Scope Accuracy Rate Trends



# Results: Lessons Learned

- While many rules used to determine scope can be aggregated to other agencies, GPO believes there will be a certain amount of customization required for each Web site.
- It has been difficult to mimic the traditionally subjective scope decision with objective rules.

# Results: Lessons Learned

- Publications in certain file formats (e.g., PDF, MS Office files), were more easily harvested accurately and in their entirety than those in HTML.
- Publications that are comprised of multiple files proved to be a challenge in that it was difficult to write rules that related the various pieces of a publication together.

# Harvesting Open Issues

- Precision and comprehensiveness
- Level of staff needed to support large amount of content from automated harvesting
- Methodologies going forward

# Next Steps

- Results of the pilot are currently being evaluated by GPO. A white paper on the final results of the pilot will be published by GPO in November.
- GPO will continue to review and compare results of its pilots with similar projects (e.g., NDIIPP Initiatives).
- GPO plans to conduct another pilot with another agency Web site that will test similar technologies and methodologies based on lessons learned.

# Next Steps (con't)

- The knowledge gained from the pilots will be leveraged in the implementation of the Harvester as a part of FDsys.
- GPO's goal is to catalog in-scope publications harvested from this pilot, starting with an investigation of the results of the automated comparison between the pilot results and GPO catalog records.
- While automated publication harvesting technology solutions are investigated and developed to improve accuracy and comprehensiveness, GPO will continue to identify and harvest publications.

# Assumptions

- GPO will use discovery, assessment, and harvesting tool(s) to identify, gather, and capture official publications from Federal Agency Web sites.
- The harvesting function will be performed either by GPO internally, an outside contractor, or a combination of the two.
- Federal Agencies will expect GPO to notify them that we are crawling/harvesting publications from their Web sites.

# Assumptions

- The harvester will be implemented in conjunction with GPO's Future Digital System (FDsys).
- The harvesting function will retrieve content and metadata necessary to create a package for ingest into FDsys, but additional processing will be required in order to complete the package for ingest into FDsys.
- Harvesting activities will follow industry best practices to ensure that GPO and target servers are not put at risk in terms of security and bandwidth.

# Discussion Questions

- Are the assumptions correct with respect to Web Harvesting?
- A harvester can be configured to harvest *only* in-scope publications or *mostly* in-scope publications including some out-of-scope publications. Given the results of the pilot, is the existing methodology sufficient to continue harvesting?

# Discussion Questions

- What other avenues regarding automated Web Harvesting should GPO be exploring in the future?
- Do you have suggestions for which agency Web site(s) would be best to focus on in a future pilot?