eotcd

Challenges in Web Archiving
UNT Perspective

NDIIPP – July 21, 2010

Broad Challenges

- Make Web archives more usable for libraries
  - Tools for collection builders
  - Bringing the Web archive to the collection builders
- Build digital library collections from Web content
  - Identification of key content within archive
  - Move beyond the needle in haystack approach of selection
- Counting and reporting
  - Understanding how Web archives should fit into traditional library metrics

EOT Archiving Project

- Who
  - Library of Congress, the GPO, the Internet Archive (IA), the University of North Texas (UNT) Libraries, and the California Digital Library (CDL)
- What
  - Snapshot of the federal government's public Web presence
- When
  - Before & after the 2009 change in administrations
- How
  - Nomination Tool: Websites
  - Website Harvests: IA, UNT, & CDL
  - Harvest Consolidation: Library of Congress

EOTCD Project

- EOT Archive Classification
  - Objective: Classify materials in accord with the Superintendent of Documents (SuDocs) Classification Numbering System
  - Outcome: Enable librarians to utilize existing selection practices to identify materials in the EOT Archive
- Web Archive Metrics
  - Objective: Identify a set of metrics for materials in Web archives
  - Outcome: Enable characterization of materials in Web archives in units of measurement more familiar to libraries and their administrations

Research Questions

1. How effective is the organization of large-scale unstructured Web archives using a pre-defined classification system, the SuDocs classification numbering system, as evaluated by government information librarians?

2. What measurable units for the materials in Web archives best support management acquisition decisions in libraries?

Next Steps

- Create a process for understanding Web archives
  - Large historical archives
  - Enable collection librarians to make decisions involving archive content
  - Workflows for moving content into other areas of the library