NORTH TEXAS

eotcd

Classification of the End-of-Term Archive: Extending Collection Development Practices to Web Archives

FDLP Conference – October 18, 2010 – Wash DC

EOTCD Project

- Classification of the End-of-Term (EOT) Archive:
  - Extending Collection Development Practices to Web Archives
- IMLS Funded
  - December 2009 – November 2011
  - Partner: Internet Archive
- Advisory Board – Representatives of other EOT Project Institutions: Internet Archive, California Digital Library, Library of Congress

EOTCD Project

Subject Matter Experts:

George Barnum – Government Printing Office
Laurie Hall – Government Printing Office
Robin Haun-Mohamed – Government Printing Office
Kevin McClure - Chicago-Kent College of Law
Michele McKnelly – University of Wisconsin, River Falls
John Phillips – Oklahoma State University
Mary Prophet – Denison University
Suzanne Sears – University of North Texas
John Stevenson – University of Delaware
Geoffrey Swindells – Northwestern University

Project Objectives

- EOT Archive Classification
  - Objective: Classify materials in accord with 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

EOT Archive 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
  - Entirety 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

EOT Web Archive: Domains

Largest Domains	# URIs	# Unique Subdomains
gov	137,780,023	14,338
com	7,805,205	57,873
org	5,107,552	29,798
mil	3,554,956	1,677
edu	3,551,845	13,856

Total # URIs: 160,156,233

EOT Web Archive: File Types

Largest Mimetypes	# Files
text/html	105,590,929
image/jpeg	13,665,196
image/gif	13,031,046
application/pdf	10,320,163

Web Archive: Problem Statements

- Current discovery methods have constraints
  - Searches commonly use URL and/or date range
  - Fulltext searches do not scale sufficiently
  - PROBLEM:
    Difficult for librarians to identify and select materials in accord with collection development policies

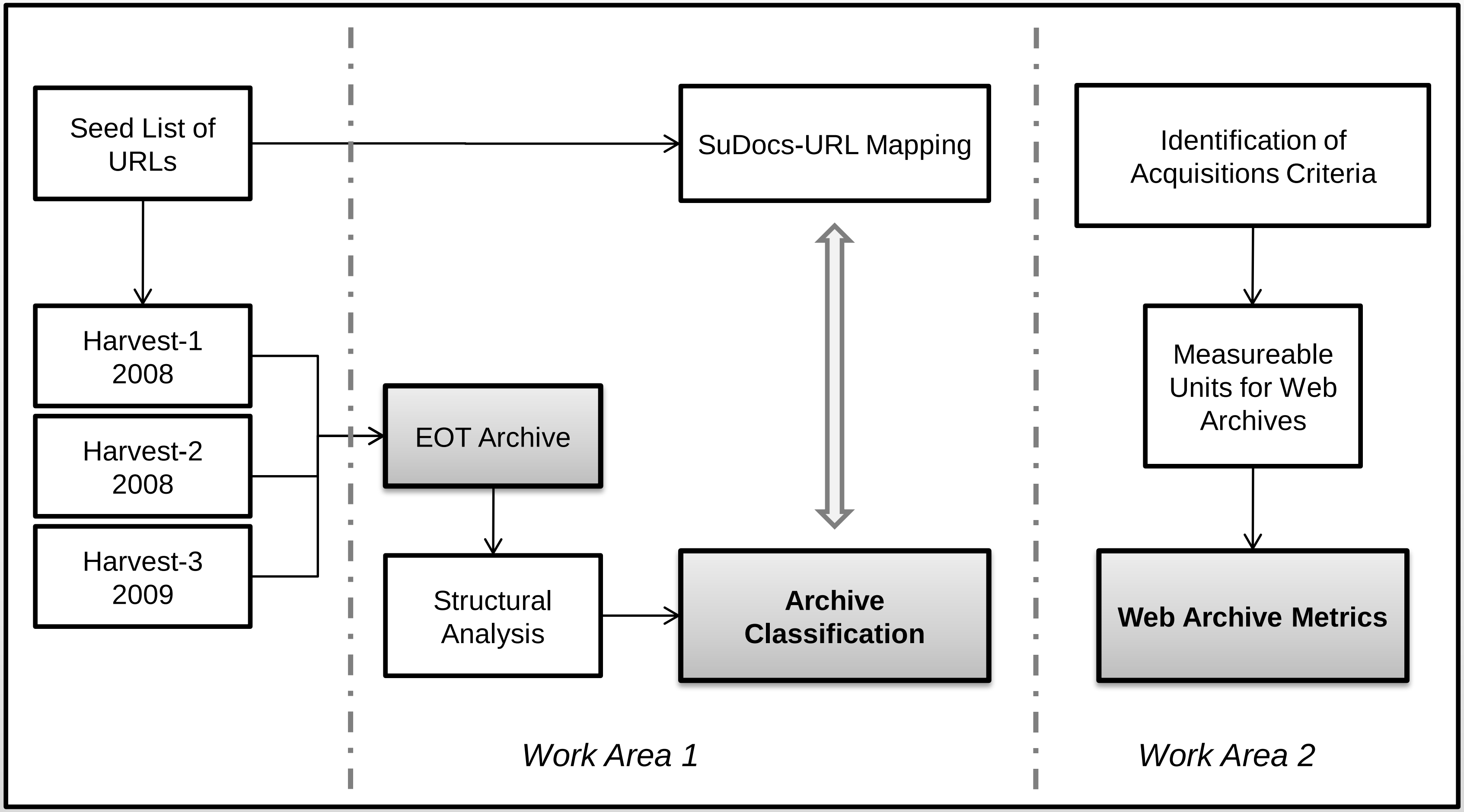
- Common metrics for materials in Web archives do not exist
  - PROBLEM:
    Difficult for librarians to communicate the scope and value of these materials to administrators

Research Questions

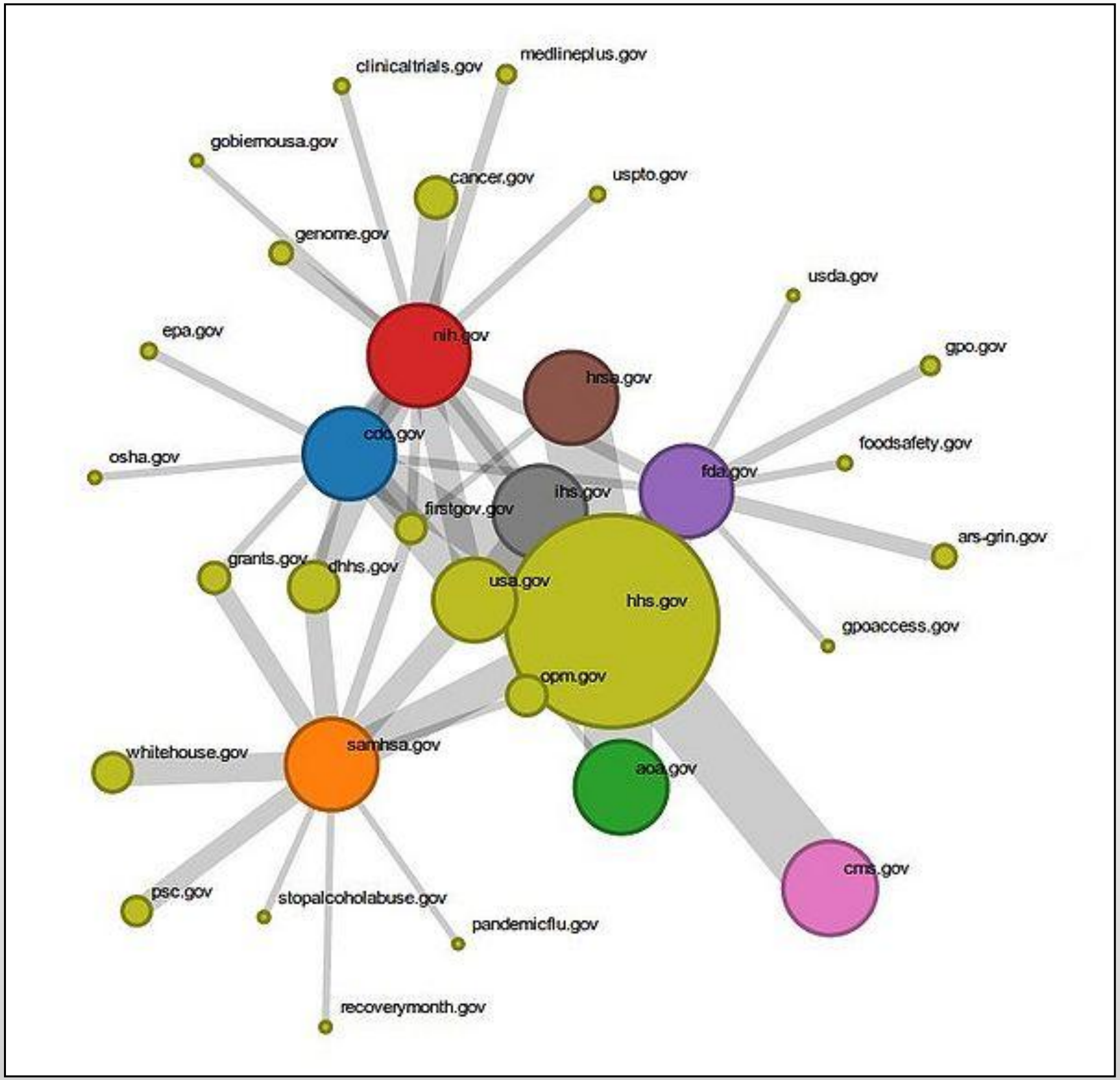
1. How effective is the organization of large-scale unstructured Web archives using a pre-defined classification system, the SuDocs classification numbering system, as evaluated by government information librarians?

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

Project Work Areas



Archive Structural Analysis



Visualization of Web Links

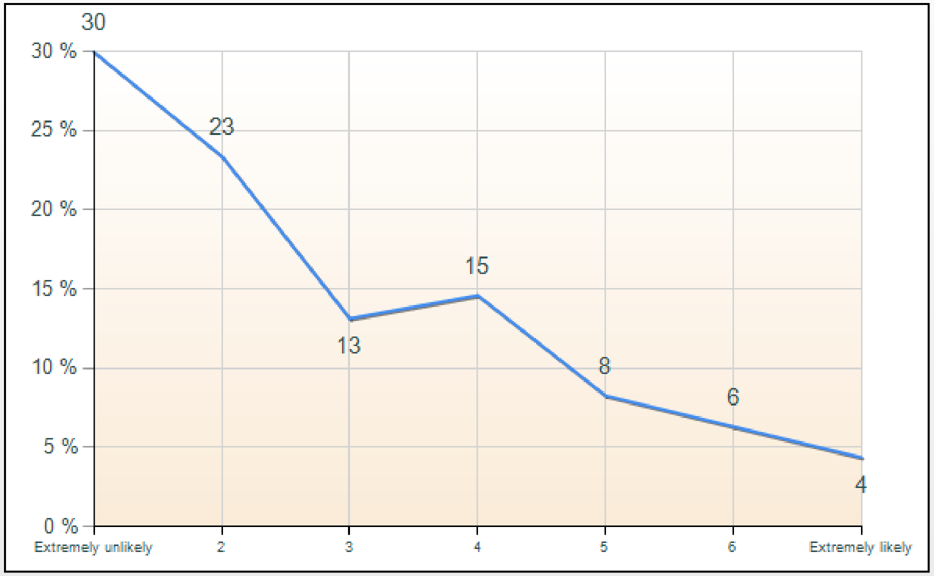
Health & Human Services Known Sub-agencies:

1. cms.gov
2. aoa.gov
3. hrsa.gov
4. cdc.gov
5. samhsa.gov
6. nih.gov
7. fda.gov
8. ihs.gov

Web Archive Metrics

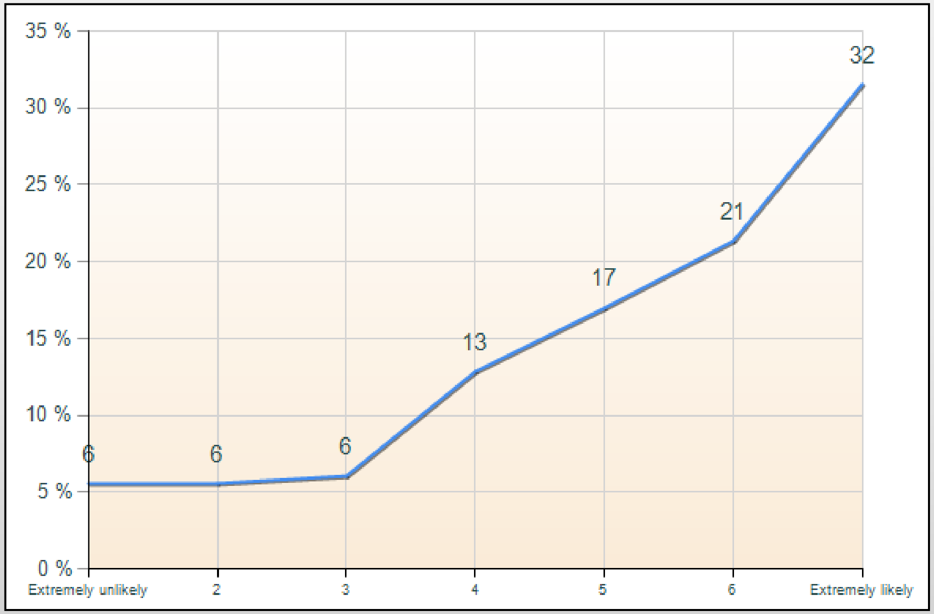
Survey of FDLP Libraries

N=416; 33% Response



% Likely to Acquire Materials

% Likely to Access Materials



Closing

Project Team:

Kathleen Murray – Senior Research Fellow & Project Manager

Mark Phillips – Technical Lead

Lauren Ko – Web Archiving Programmer

Graduate Research Assistants: Cathy Benton & Bharath Dandala

Cathy Hartman – Project Oversight

Project Website

http://research.library.unt.edu/eotcd

Thank you!