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

SME Meeting – April 25, 2010 – Buffalo, NY

Agenda

11:30 AM	Lunch
12:30 PM	Project Overview
1:45 PM	Break
2:00 PM	Focus Group Discussion
3:45 PM	Closing Remarks
4:00 PM	End

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

Project Objectives

- 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

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

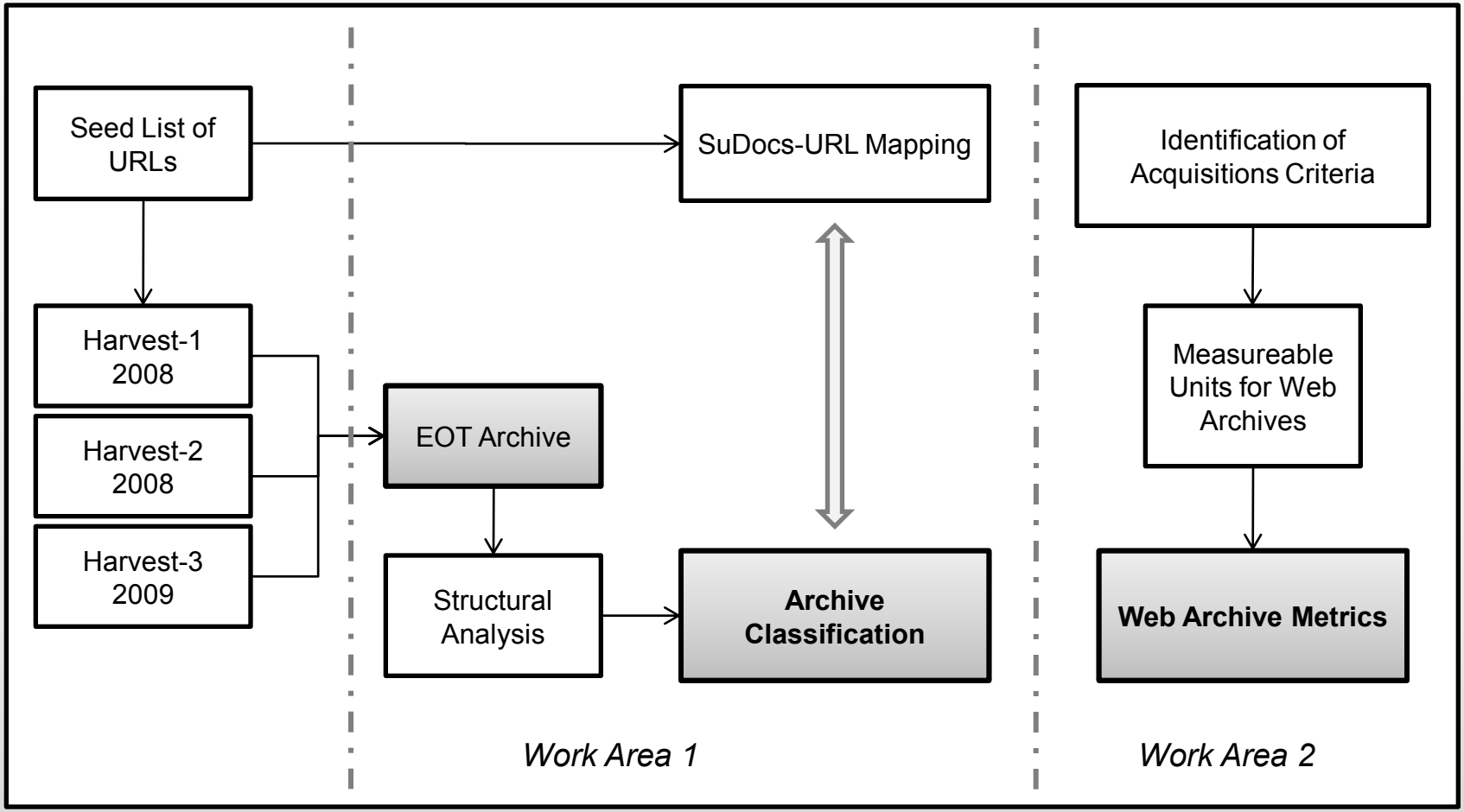
- Current discovery methods have constraints
  - Searches commonly use URL and/or date range
  - Fulltext searches do not scale sufficiently
  - PROBLEM:
    Difficult for librarians to identify and select materials in accord with collection development policies
- Common metrics for materials in Web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

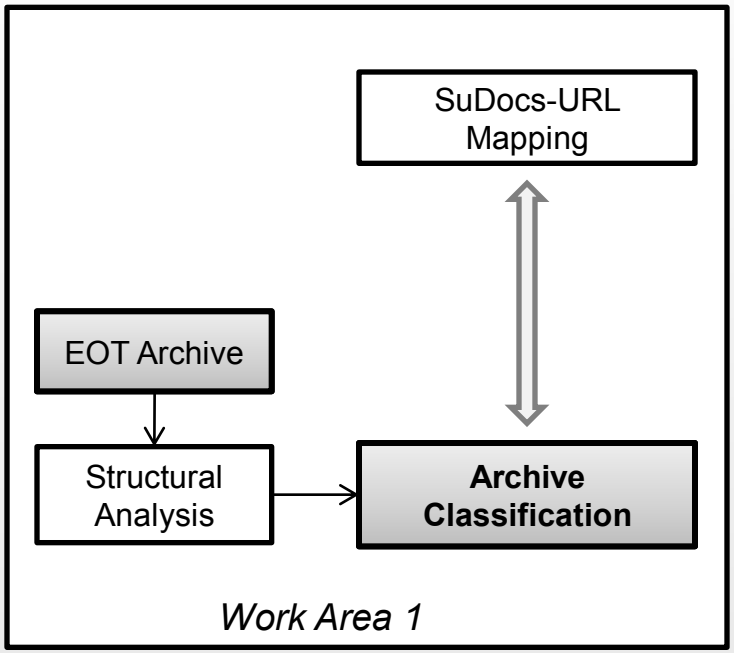
2. What measurable units for the materials in Web archives best support management acquisition decisions in libraries?

Project Work Areas



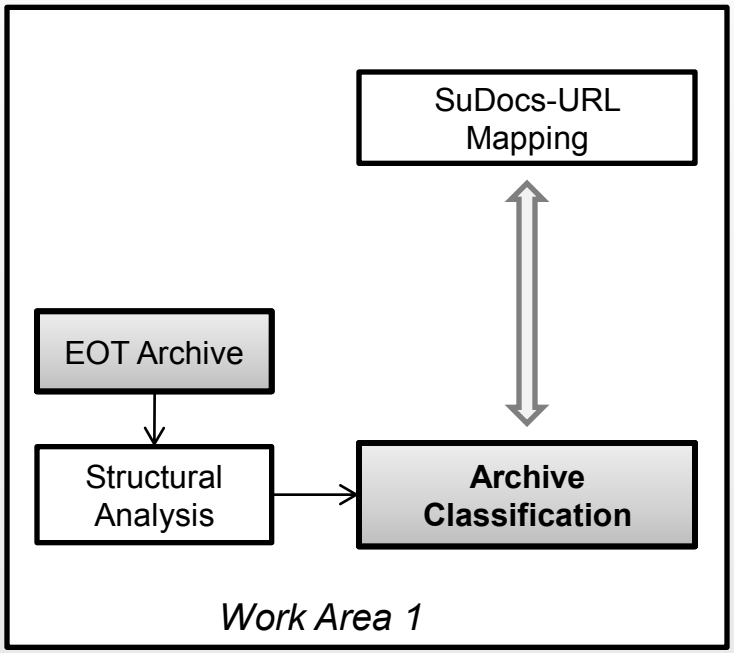
Archive Classification

- UNT: Structural Analysis
  - Link analysis tool (IA) & visualization tools
  - Identify organizational & relational structure
  - Assign SuDocs numbers to the structured Archive
- SMEs: Human Analysis
  - SuDocs to URIs in Archive
    - UNT Mapping Tool
    - Oct-Dec 2010
  - Result: SME Map



Archive Classification

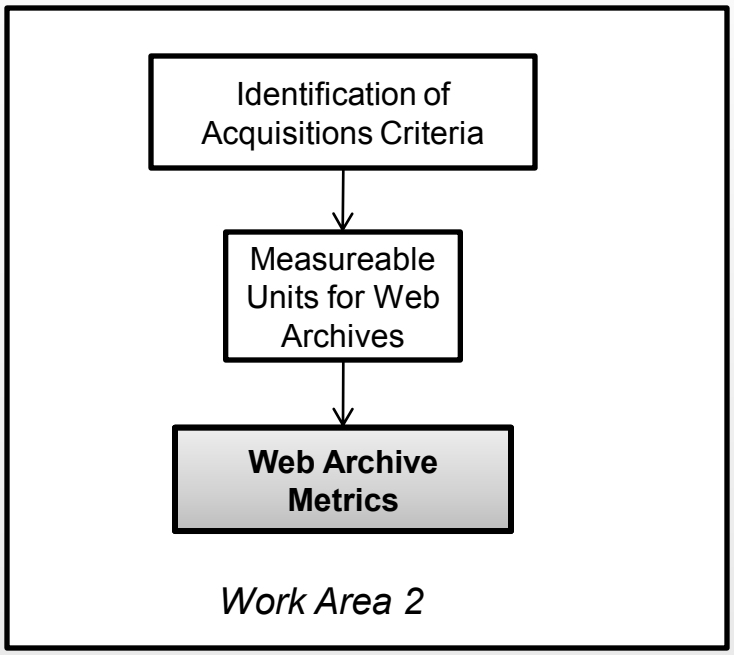
- UNT: Structural Analysis
  - Link analysis tool (IA) & visualization tools
  - Identify organizational & relational structure
  - Assign SuDocs numbers to the structured Archive
- SMEs: Human Analysis
  - SuDocs to URIs in Archive
    - UNT Mapping Tool
    - Oct-Dec 2010
  - Result: SME Map



RQ1: Effectiveness of Structural Analysis:
Compare structural analysis results to SME Map
SME Evaluation: April 2011

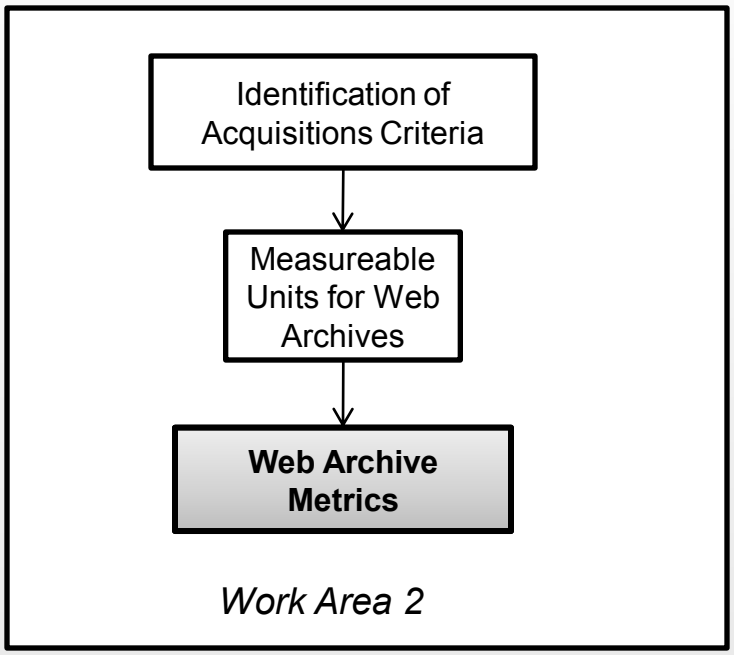
Web Archive Metrics

- SMEs: Acquisitions Criteria
  - Focus Group 1: April 2010
  - Statistical reporting
- UNT: Equivalencies
  - Translation tool
  - Archive measureable units to library measurement units
- Experimental Collections
  - SMEs: Selection Lists
  - UNT: Collections & Metrics
  - SMEs: Evaluate Results
    - Focus Group 2: October 2011



Web Archive Metrics

- SMEs: Acquisitions Criteria
  - Focus Group 1: April 2010
  - Statistical reporting
- UNT: Equivalencies
  - Translation tool
  - Archive measureable units to library measurement units
- Experimental Collections
  - SMEs: Selection Lists
  - UNT: Collections & Metrics
  - SMEs: Evaluate Results
    - Focus Group 2: October 2011



RQ2: Measureable Units for Materials in Web Archives:
Identify Common Metrics

Closing

- Project Website
  - http://research.library.unt.edu/eotcd
  - Focus group report
  - Project information
- Next SME Meetings
  - October 2010: Washington DC
  - April 2011: DLC Location
  - October 2011: Washington DC

Thanks very much for your participation!