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

SME Meeting – October 17, 2010 – Wash DC

Agenda

11:30 AM	Working Lunch - Metrics Discussion
12:30 PM	Project Update: Web Archive Metrics
1:00 PM	Break
1:15 PM	Project Update: Archive Classification
1:45 PM	Classification Exercise
3:15 PM	Closing Remarks
3:30 PM	End

Metrics Focus Group: Selection Criteria

- Broadness of applicability
  - Scope or breadth of material coverage to serve the "broadest possible group of users"
  - Promotes buy-in from multiple departments
- Usage data
  - Generally vendor provided
  - Vendor compliance with standards needed
- Appropriateness for collection
  - Particularly in regard to the degree of "completeness" needed for in a particular subject

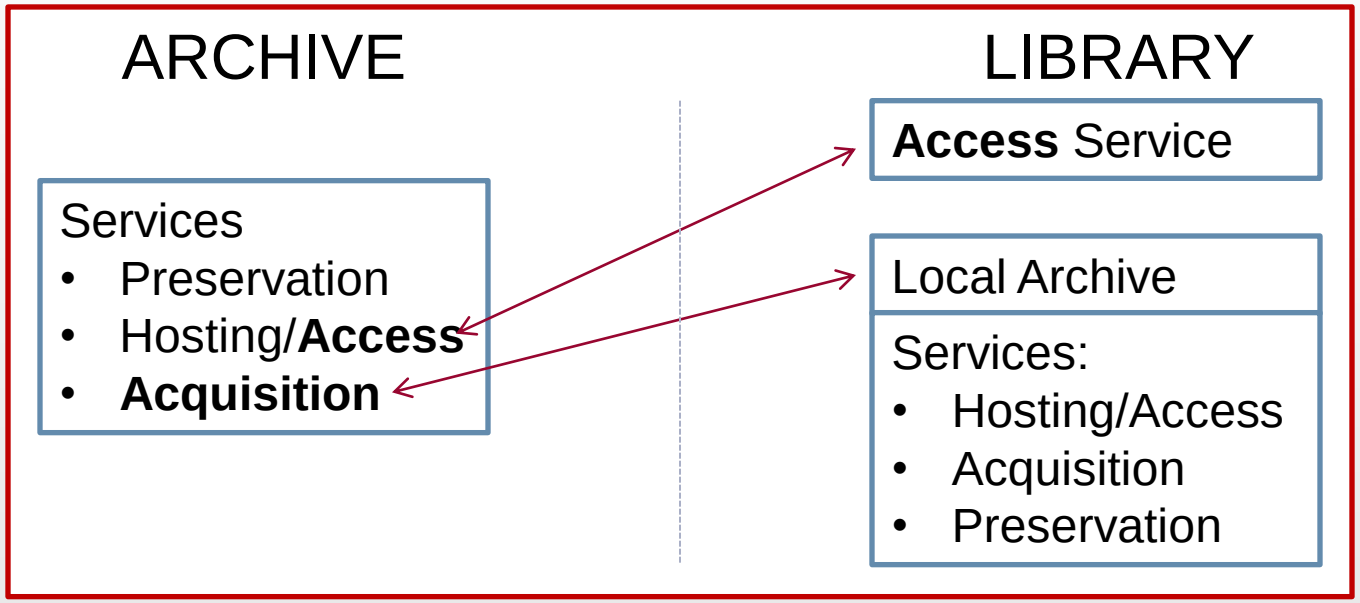
Metrics Focus Group: Selection Criteria

- Number of titles
  - A measure of the volume or amount of materials
- Unique content
  - Number of unique items in the archive, that is, materials not available elsewhere
- Duplicate content
  - The “titles” (or materials) in the existing collection that are duplicated

Metrics Focus Group

- Essential requirement for selection decisions:
  - Standard data elements for comparable material types
  - For networked electronic resources, counts based on IP addresses for:
    - Specific pages and collections *accessed*
    - Specific files/materials *retrieved*
- Metrics that drive acquisitions
  - Retention: Cost per use
  - Selection: Usage data (when available)

Metrics Focus Group: Service Models



1. Access Model
2. Acquisition Model

Metrics: Library Statistics

1. Access Model
2. Acquisition Model

- Categories
  - Scope (How much; how many)
  - Expenditures (Cost)
  - Usage (Counts)
  - Quality (Outcomes; Value)
- Authorities - Standards
  - ARL; ACRL; IPEDS
  - COUNTER: Codes of Practice
    - SUSHI: ANSI/NISO Z39.93-2007
      - Standardized Usage Harvesting Initiative
  - ISO TC46/SC8/WG9
    - *Statistics and quality issues for web archiving*

Web Archive Metrics: Perspectives

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

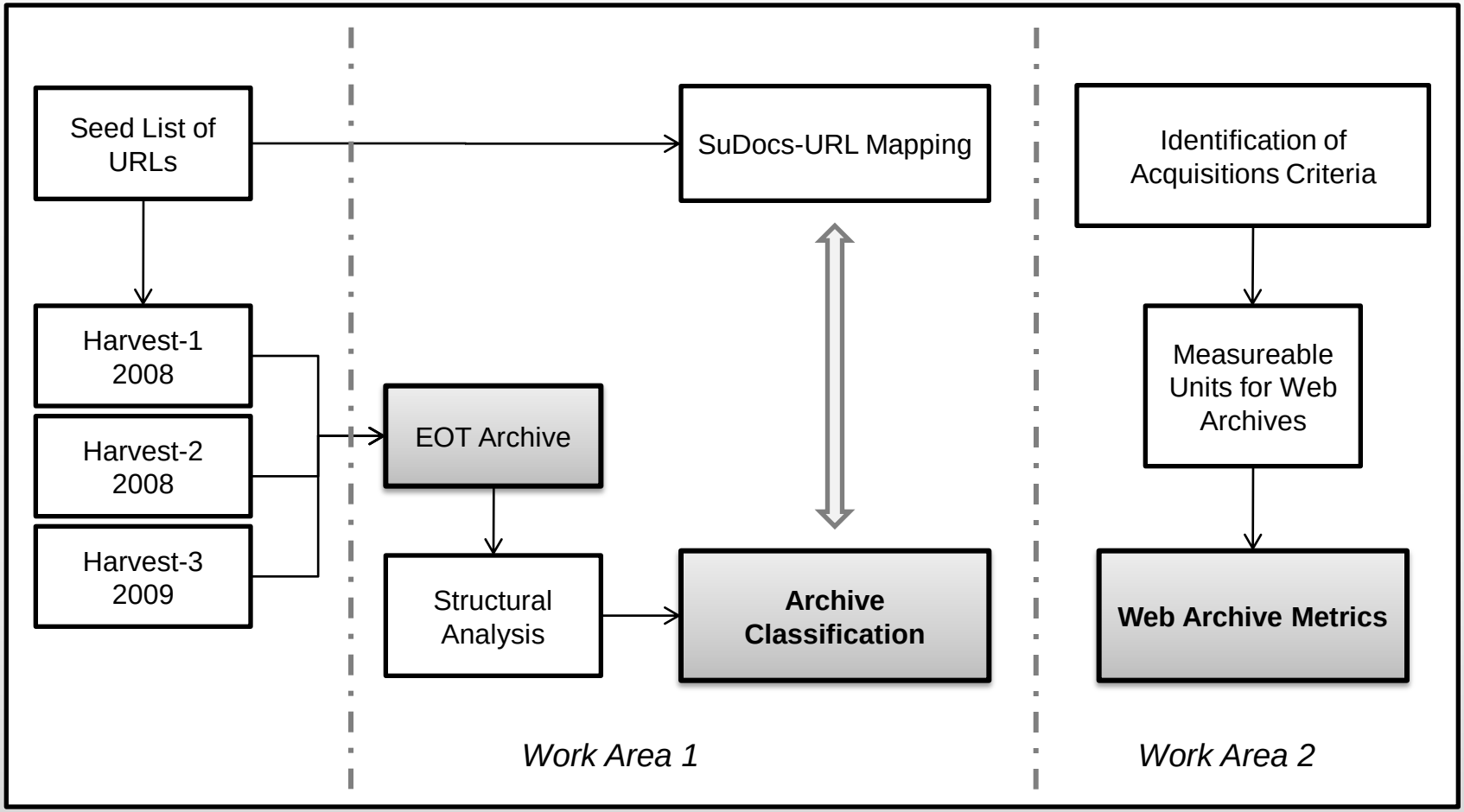
"In *Web Archive Metrics*, a draft prepared for the IIPC, Boyko distinguishes between metrics for *internet-based* aggregations and for *collection-based* aggregations of Web pages. The need for the two sets of metrics reflects different scenarios for future use of the archived Web pages. Some researchers will want to study the Web as a network, analyzing patterns of links and changes over time. Others will want to locate materials of a particular type (e.g., blogs) or pages devoted to a particular topic." - *Library of Congress*

Web Archive Metrics: Perspectives

- ARL New Measures Initiative
  - StatsQUAL®: A Gateway to Library Assessment Tools
  - "New measures that address issues of library service quality, electronic resource usage and value, and outcomes assessment."

- COUNTER
  - ARL New Measures Initiative: "set up in response to the following two needs:
    - increasing demand for libraries to demonstrate outcomes/impacts in areas important to the institution, and
    - increasing pressure to maximize use of resources."

Project Status



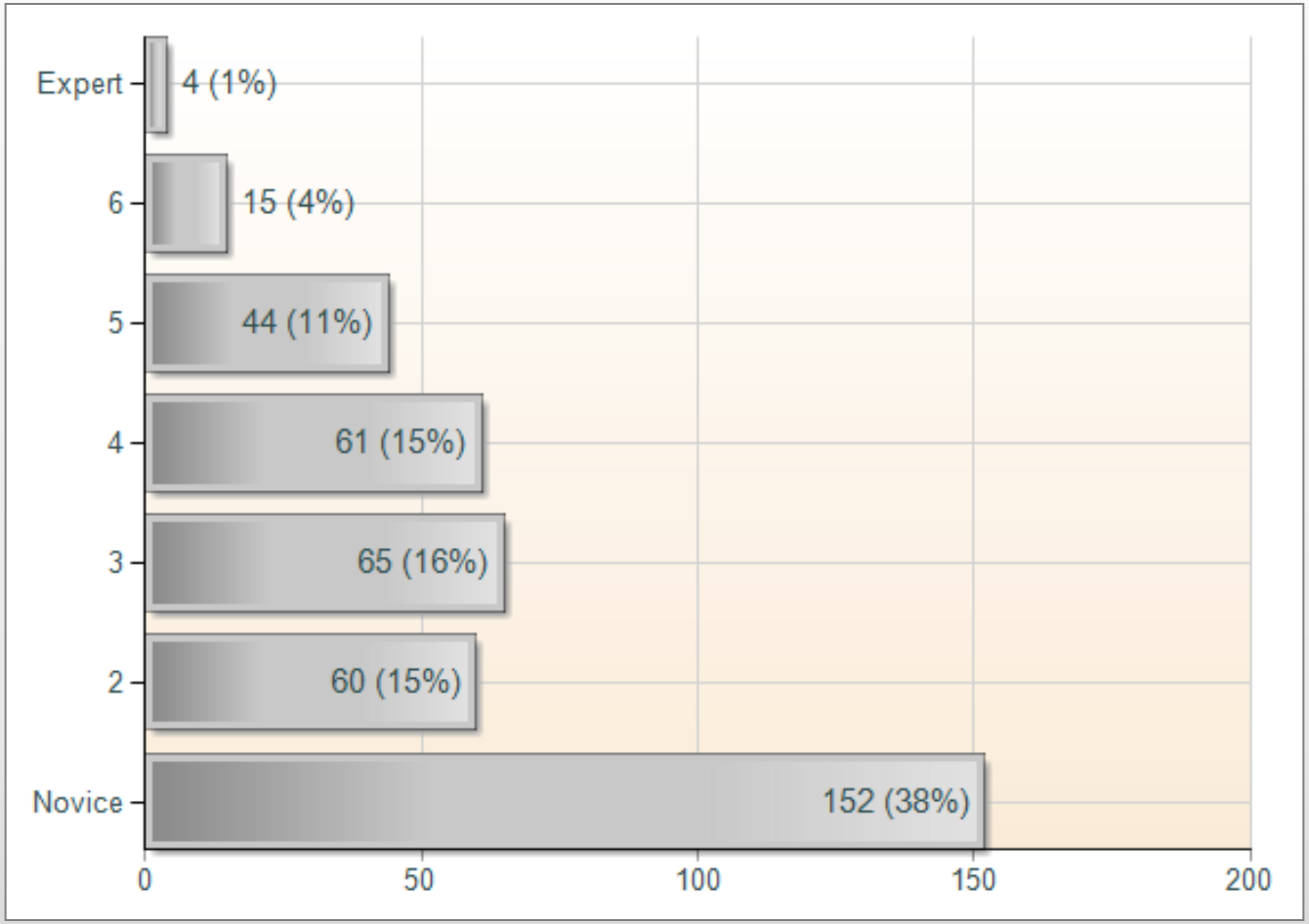
Work Area: Web Archive Metrics

- FDLP 2009 Biennial Survey: Q18b Data
  - *Are you interested in receiving digital files on deposit?*
  - *Have you discussed this with your library director or dean?*
  - *Is there administrative support for receiving digital files on deposit?*
- eotcd Project Survey of FDLP Libraries
  - Assess interest in access to v. acquisition of materials from web archives
  - Assess libraries' capabilities to support acquisition:
    - Hosting and access
    - Long-term preservation

FDLP 2009 Biennial Survey: Q18b Results

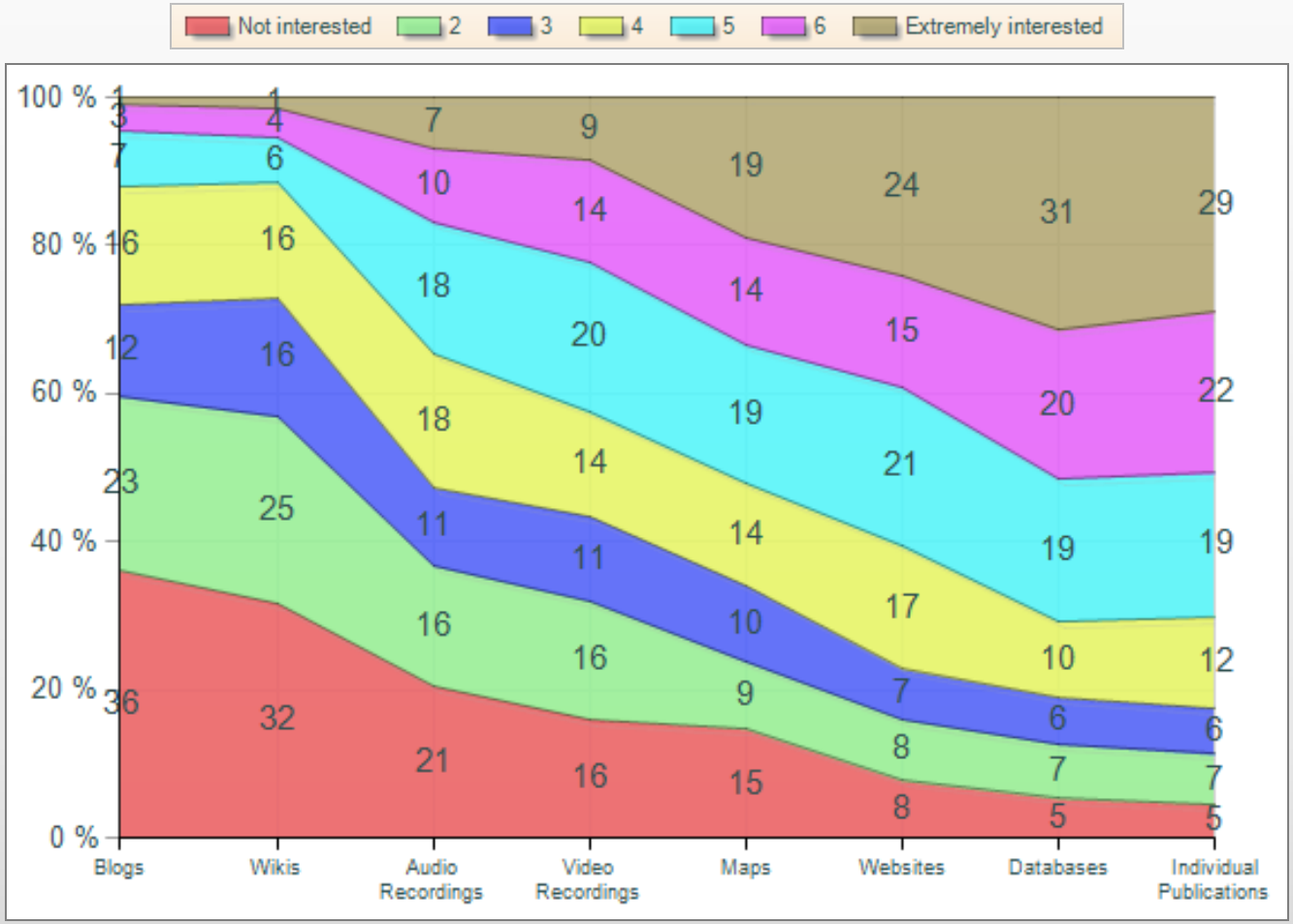
Q 18b-1 Are you interested in receiving digital files on deposit?		
YES: # - %	417	37%
Q 18b-1&2: Are interested and have discussed with library director/dean		
YES: # - %	280	25%
Q 18b-1&3: Are interested and there is administrative support		
YES: # - %	337	30%
Q 18b-1,2,&3: Are interested, have discussed with library director/dean, and there is administrative support		
YES: # - %	249	22%

Survey of FDLP Libraries (*N*=416; 33% Response Rate)



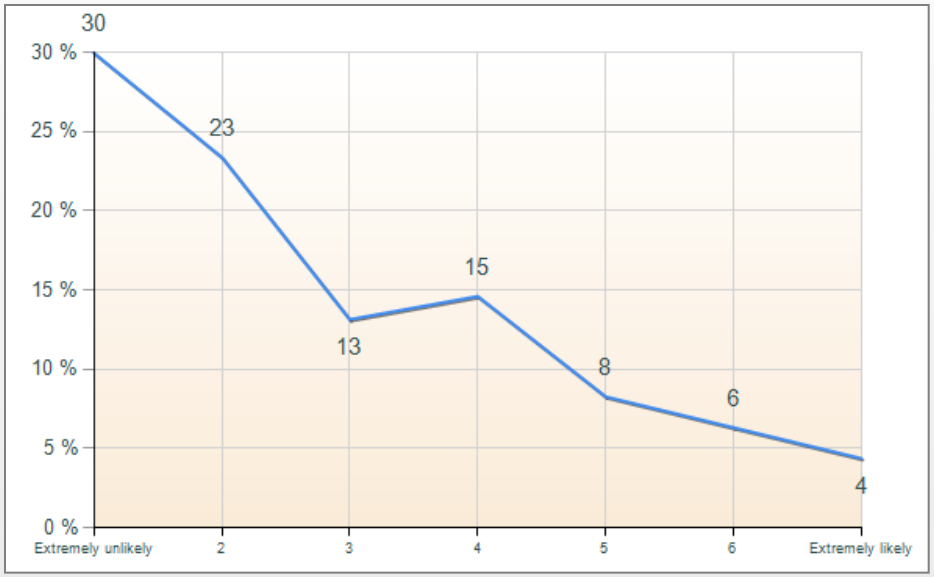
Experience with Web Archives

Survey of FDLP Libraries



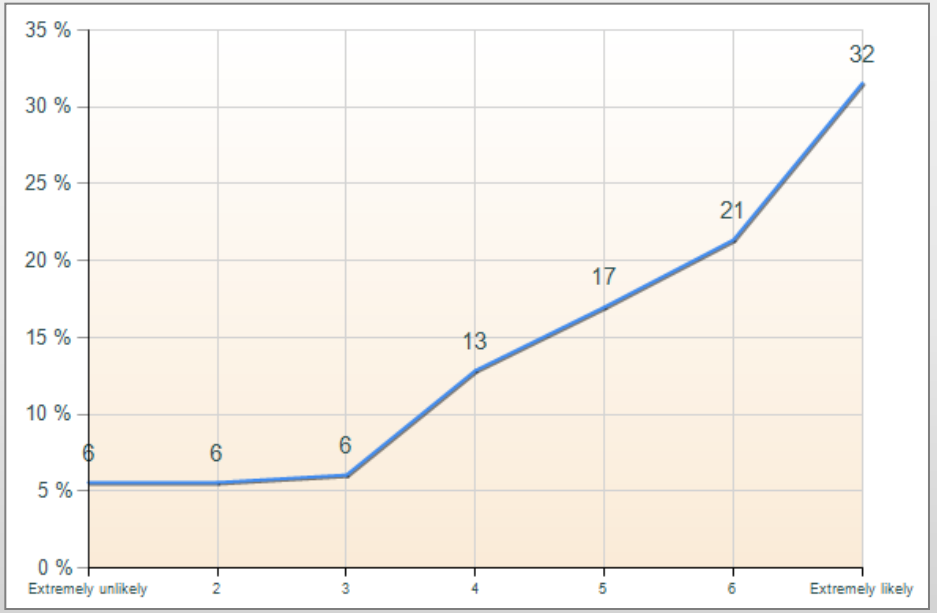
Interest in Materials by Type (%)

Survey of FDLP Libraries

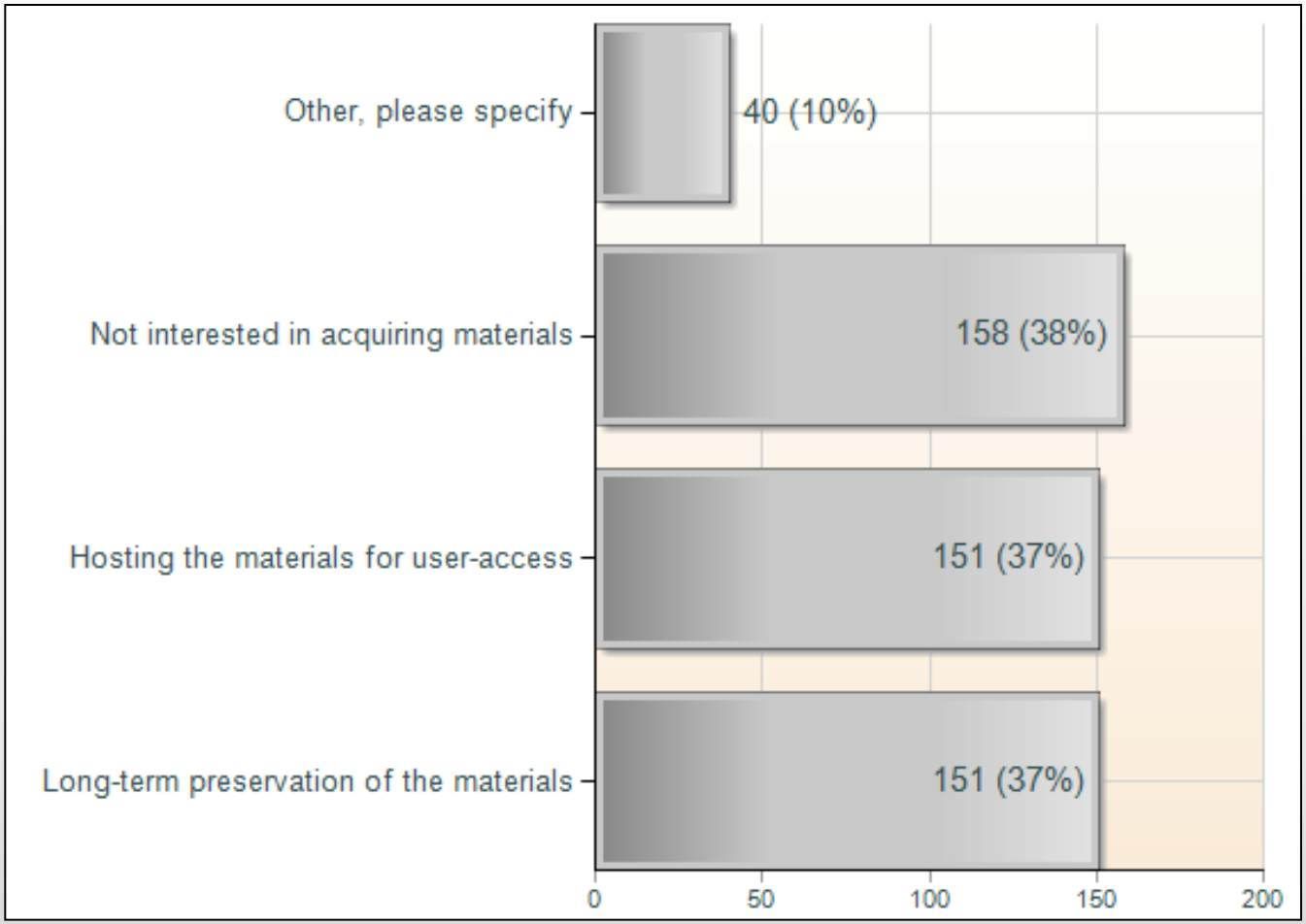


Likely to Acquire Materials (%)

Likely to Access Materials (%)

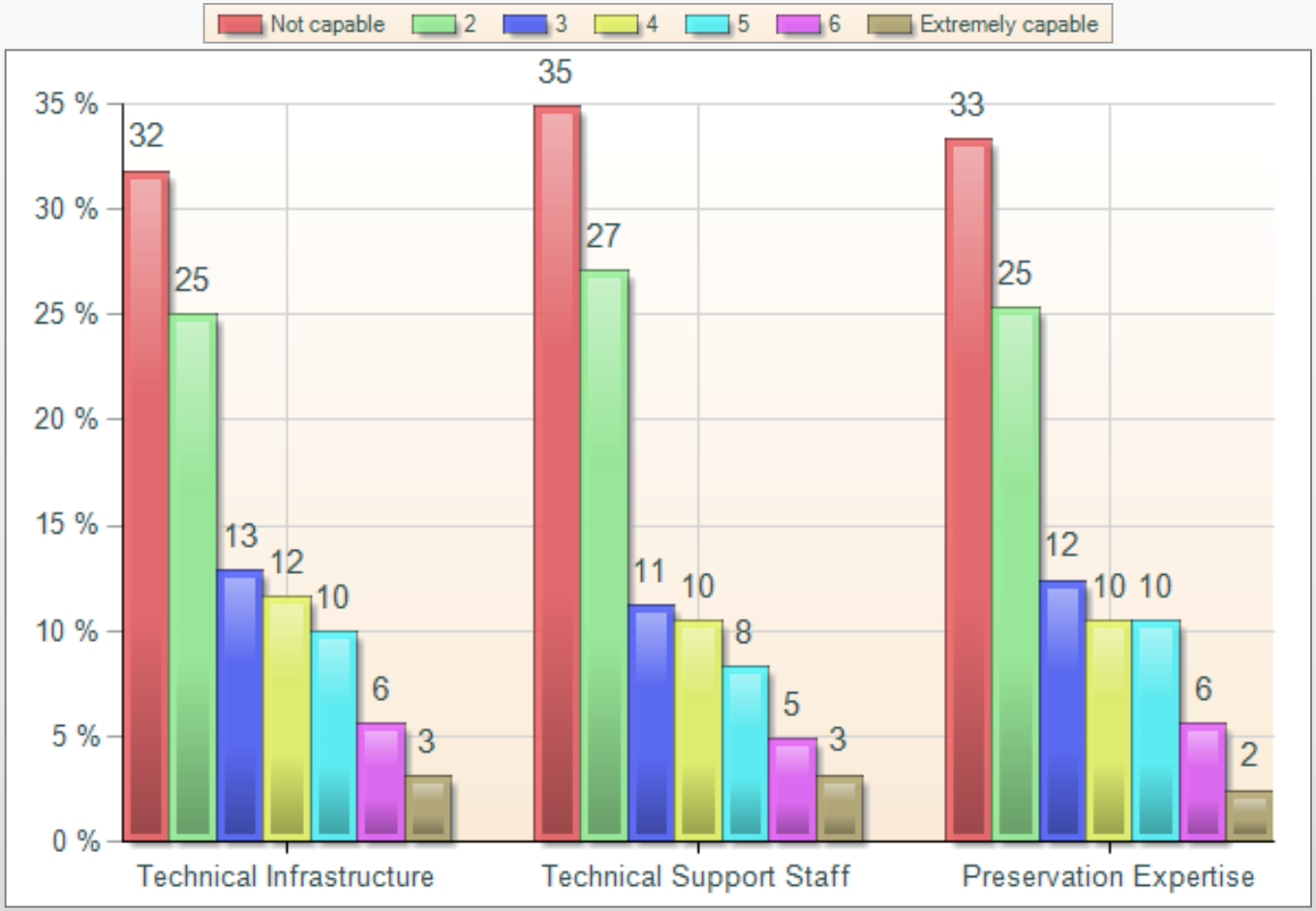


Survey of FDLP Libraries



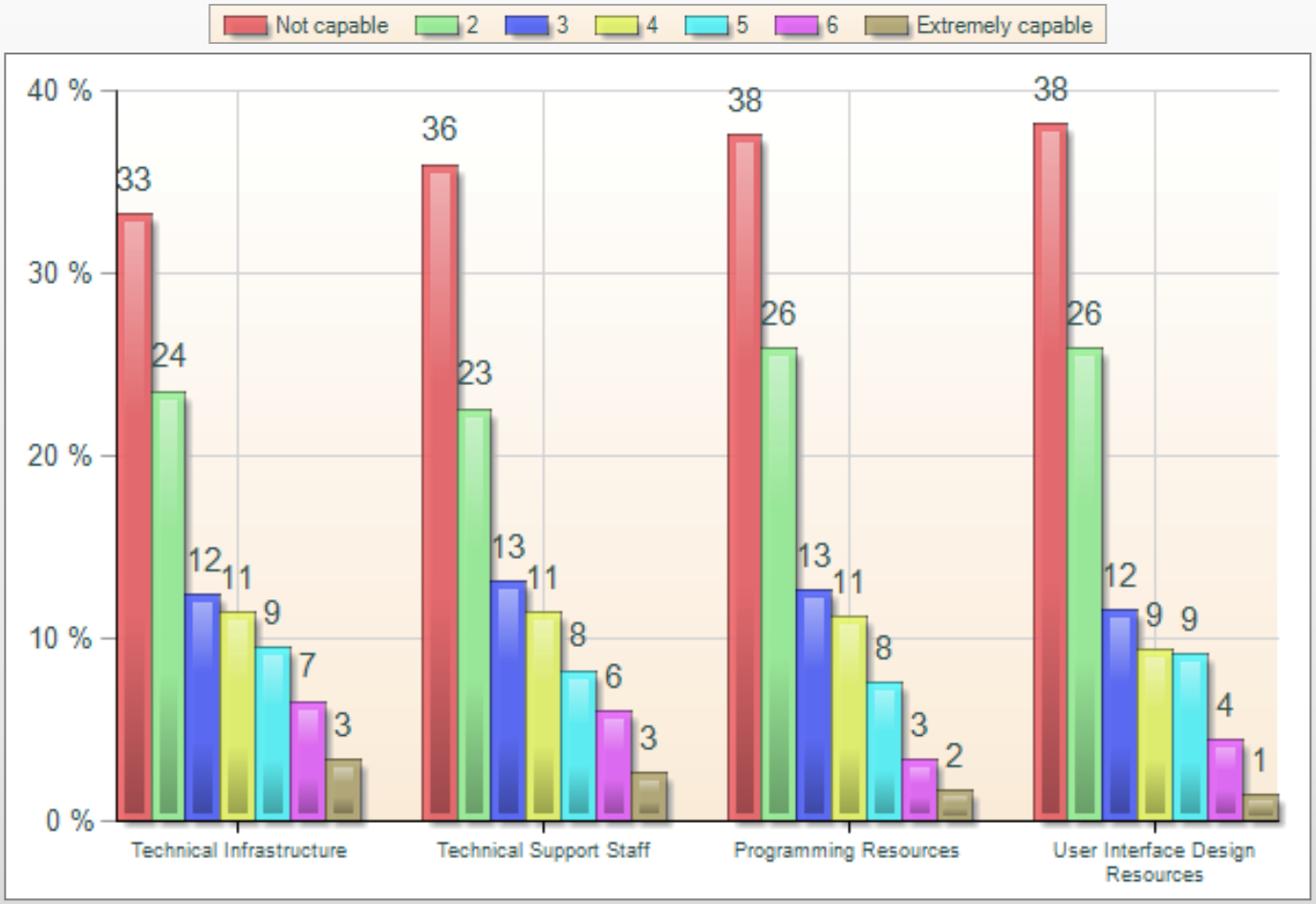
Motivation for Acquisition

Survey of FDLP Libraries



Capability to Support Long-Term Preservation (%)

Survey of FDLP Libraries



Capability to Support Hosting & User Access (%)

BREAK

Work Area: Archive Classification

- Sampling the EOT archive
  - Structural Analysis
  - Classification

- Structural analysis
  - Visualizations

Sampling the EOT Archive

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

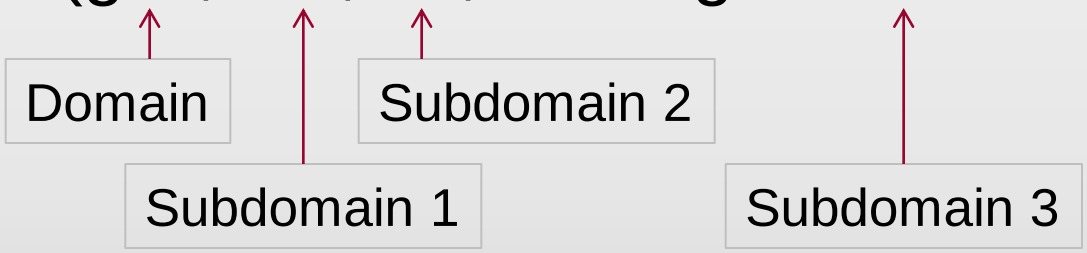
Reduced Unique Subdomains to 16,015

Sampling the EOT Archive

SURTS: Reordering URIs by domain structure

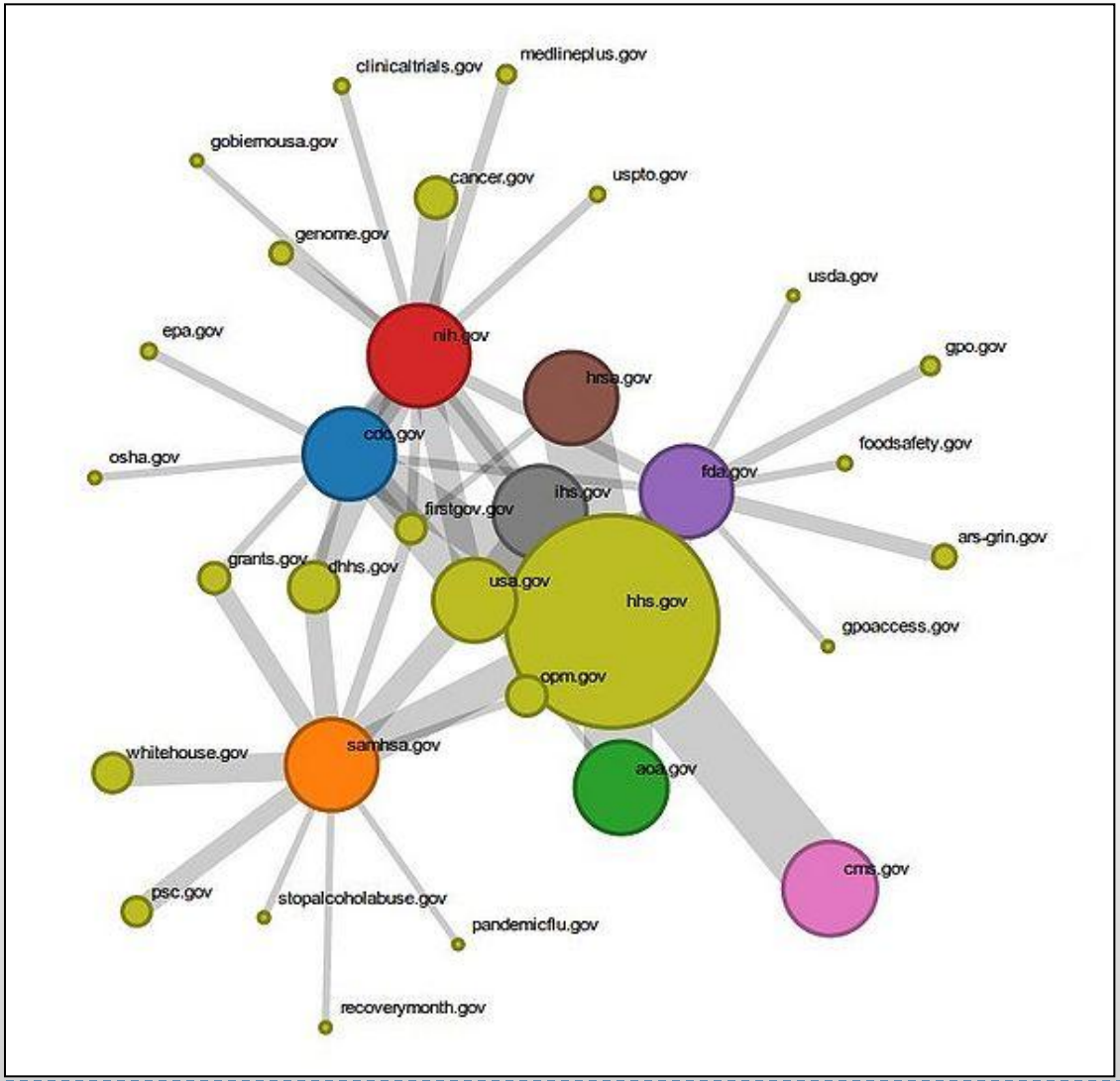
Example URI:
http://marriagecalculator.acf.hhs.gov/marriage/

SURT:
http://(gov,hhs,acf,marriagecalculator,)



Unique Subdomains 1st Level = 1,151

Archive Structural Analysis

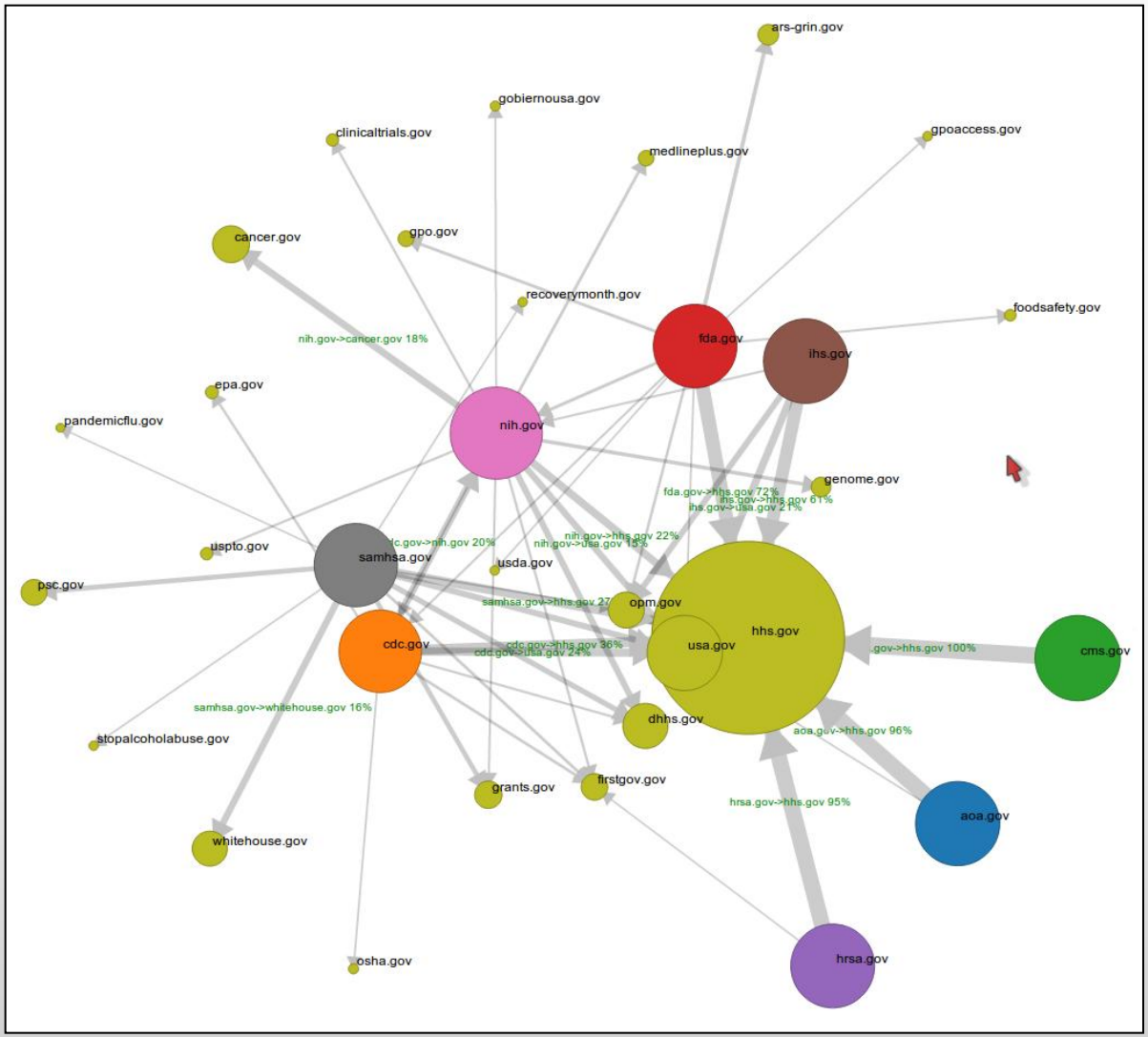


Visualization of Web Links

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

Archive Structural Analysis



Visualization of Directional Outlinks

Health & Human Services Known Sub-agencies:

- Static View
- Interactive View

SME Classification Exercise

- Key Points
  - Completion Date: Friday, November 19, 2010
  - Two people will classify every URI
    - Resolution of differences
  - Email next week
    - Username/password
    - URL for the Classification Application

- Exercise
  - Enter *keywords*
  - *Select* a SuDocs number from list
  - Enter *multiple* agency authors as appropriate

Closing

- Project Website
  - http://research.library.unt.edu/eotcd
    - Reports & updates
    - Work in progress
- Expense Reports
- Next SME Meetings
  - April 2011: DLC Location
  - October 2011: Washington DC

Thanks very much for your participation!