Curation of the End-of-Term Web Archive

Kathleen Murray – University of North Texas Libraries

Federal Depository Library Conference – October 2011 – DC

Topics

- Background
  - EOT Web Harvest Project
  - EOTCD Project
- Classification
  - SMEs: SuDocs
  - Link Analysis: Web graph
    - Cluster Analysis
  - SMEs : Cluster Tagging
- Overall Findings
  - Clusters: SuDoc Classification & Tagging
- Metrics
- Closing

Background: EOT Web Harvest 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 Archive: 16 Terabytes

Largest Domains	# URLs	# Unique Subdomains	
gov	137,847,822	14,339	→ gpo.gov
com	7,809,711	57,873	
org	5,108,645	29,798	
mil	3,555,425	1,677	→ army.mil
edu	3,552,509	13,856	

EOT Archive: File Formats

File Format	# URLs	
Text	109,498,363	→ html, plain
Image	29,140,868	
Text-like	11,234,522	→ pdf, msword
Computer file	3,472,193	
Dataset	908,339	
Video	318,498	
Audio	198,349	

EOTCD Project

- EOTCD Project
  - Classification of the End-of-Term (EOT) Archive: Extending Collection Development Practices to Web Archives
  - IMLS Funded (IMLS Award LG-06-09-0174-09)
  - December 2009 – November 2011
  - Partners: UNT Libraries & Internet Archive
- Subject Matter Experts (SMEs)
  - 12 Government information professionals
- Advisory Board
  - US members of IIPC from End-of-Term Web Harvest Project

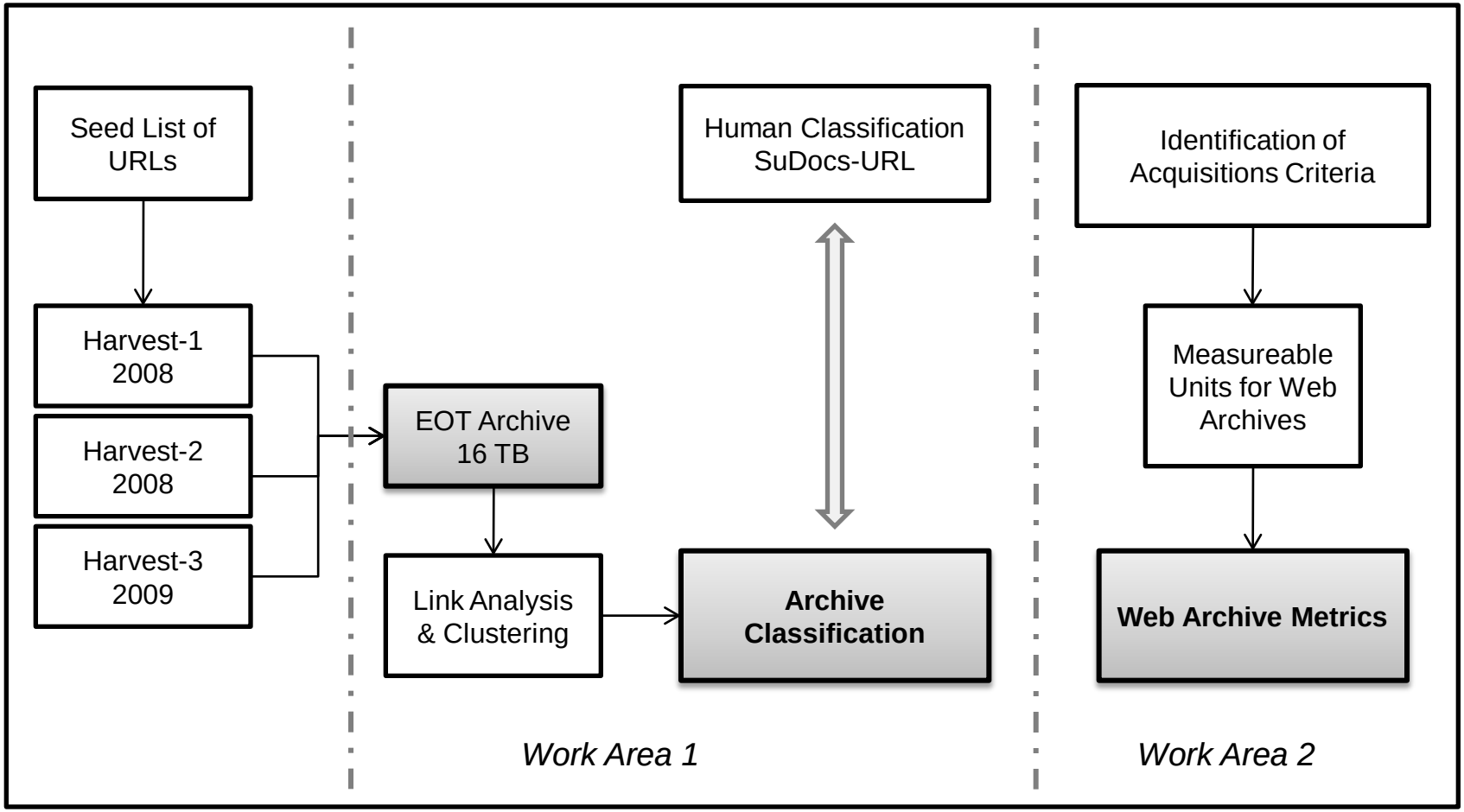
Background: Problem Statements

- Selection of Materials
  - WARC files (ISO 28500)
    - Specifies formats needed for storage, management, and exchange of data objects (or resources)
    - Applications required to discover and render resources
  - Wayback access
    - Foreknowledge of a resource's URL often required
  - The absence of descriptive metadata or classification schemes thwarts discovery & access
- Metrics
  - Acquisition & retention decisions require standard metrics which are not available

Background: EOTCD 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

Background: EOTCD Work Areas



CLASSIFICATION

Classification: Challenges

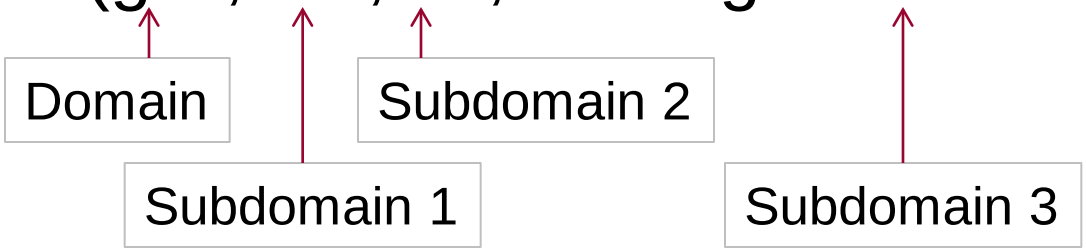
Largest Domains	# URLs	# Unique Subdomains
gov	137,847,822	14,339
com	7,809,711	57,873
org	5,108,645	29,798
mil	3,555,425	1,677
edu	3,552,509	13,856

Reduced Unique Subdomains to 16,016

Classification: Managing the Size

SURTS: Reordering URLs by domain structure

Example URL:
http://marriagecalculator.acf.hhs.gov/marriage/
SURT:
http://(gov,hhs,acf,marriagecalculator,)

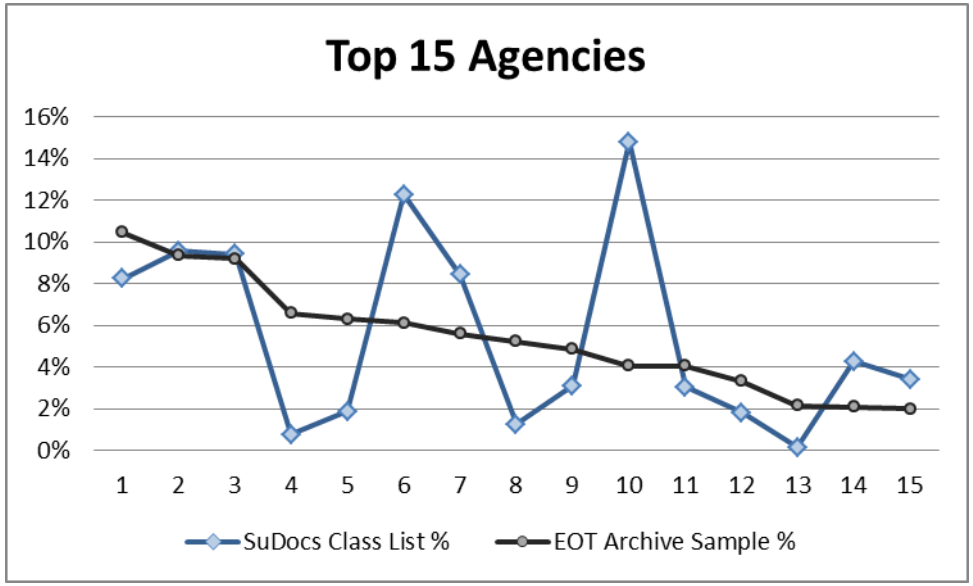


Unique Subdomains 1st Level = 1,647
After validation = 1,151 Subdomains

Human Classification

- SuDocs Classification System
- 10 SMEs classified 1,151 URLs (230/SME)
  - 70% agreement (*n* = 808); 30% disagreement (*n* = 343)
  - Unable to classify: 18 - in scope; 36 - out of scope
- 3 arbitrators classified 343 URLs
  - Assigned SuDocs authors to 286 URLs
  - Unable to classify: 42 - in scope; 15 - out of scope
- Final result:
  - Assigned SuDocs authors to 1,040 subdomains
  - 1,111 authors (1,040 + 71 multiply authored sites)

Findings: Federal Agency Representation



- 15 Agencies Represent:
  - 81% of authors in EOT Archive sample
  - 82% authors in SuDocs class list
- * 2 Agencies: Near identical percentages
  - D and HE
- ** 3 Agencies: Differ by 5% or more
  - GS, C, A

	Agency
1	Congress
*2	Defense Department
*3	Health and Human Services Department
**4	General Services Administration
5	Treasury Department
**6	Commerce Department
7	Interior Department
8	Executive Office of the President
9	Energy Department
**10	Agriculture Department
11	Justice Department
12	Homeland Security
13	President of the United States
14	Transportation Department
15	Labor Department

Findings: Feedback

- SuDocs Classification System
  - Overall, it worked well to classify Websites
  - Lacks sufficient granularity for subordinate offices and agencies
  - Forced to classify at high level

- Major Classification Challenges
  - Determining primary author among multiple authors

Link Analysis

- Subdomains
  - 1,151 1st level subdomains within .gov & .mil domains
  - Multiple URLs per subdomain
- Web graph
  - Identified # of outlinks and inlinks for each URL
- A number of cluster analysis algorithms explored
  - Best result: Agglomerative Hierarchical Clustering

Cluster Analysis

- Set limit on number of clusters to identify
  - First analysis: Set of 55 clusters
  - Second analysis: Set of 75 clusters

Cluster Set 55 - #24
7 Subdomains
• fdic.gov • fdicconnect.gov • fdicig.gov • fdicoig.gov • fdicseguro.gov • myfdicinsurance.gov • egrpra.gov

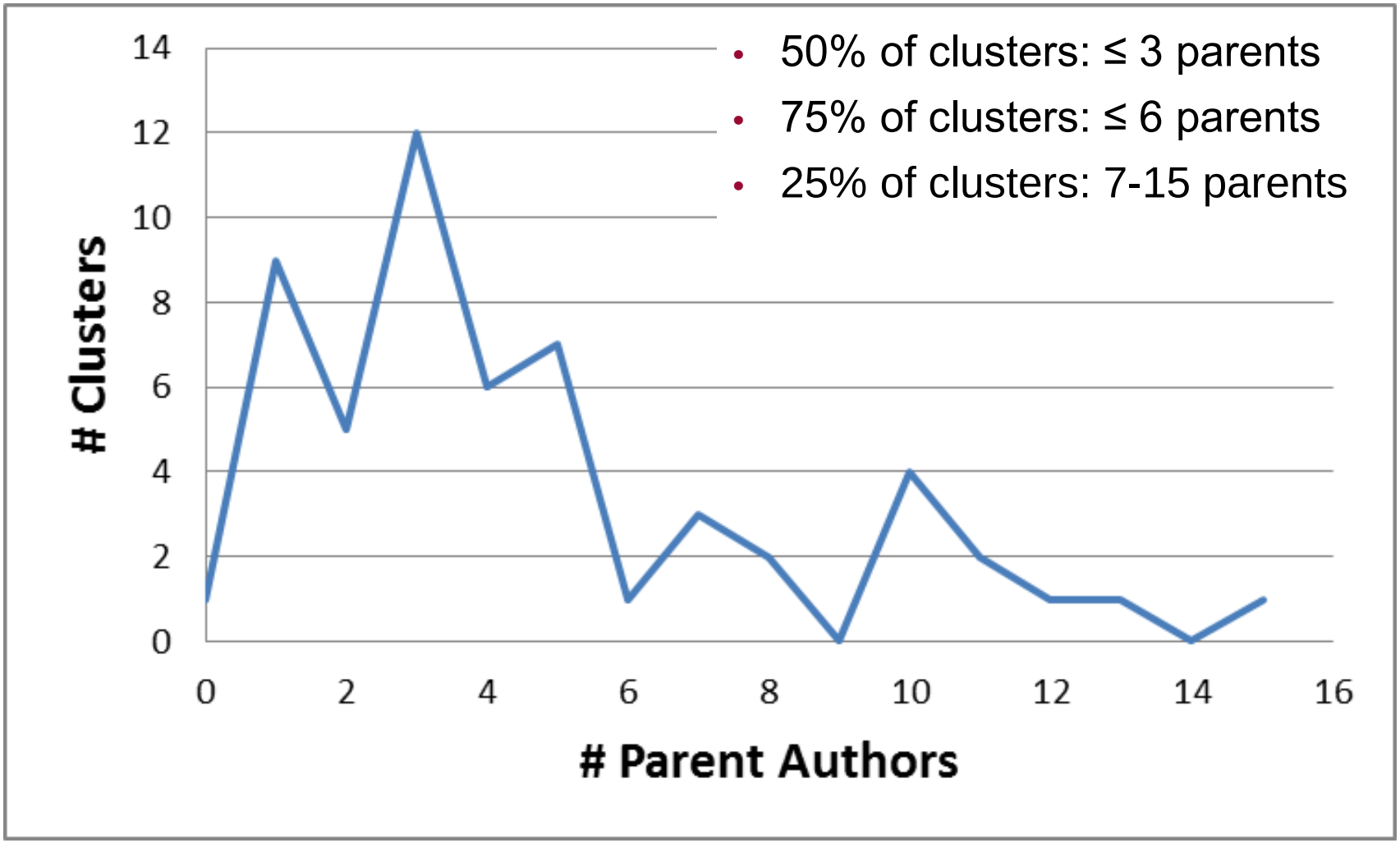
Cluster Set 75 - #63
3 Subdomains
• usccr.gov • fmcs.gov • adr.gov

Cluster Analysis: Findings

- EOT Archive reflects the variances in government agency authors
  - Size; number & size of sub-agencies; amount published
- Evaluation: Clustering in geometric space is problematic when Web graph is highly linked and its density is highly variable throughout

NOTE: Clusters on project wiki: http://research.library.unt.edu/eotcd/wiki/Clusters

Subdomain Classification: 55 Clusters



Conclusions

- Involving SMEs in classifying a reasonable sample of a domain-specific Web archive might enable their expertise to be leveraged to:
  - Improve cluster analysis
  - Increase the relevance of search results
- Cluster analysis suggests topical groupings across agency authors
  - Often with 1-2 dominant agency authors
  - Implication for search results:
    - Suggest possible related sites of interest in support of cross-agency subject-related content

Cluster Tagging

Cluster Tagging Exercise

- Total of 130 clusters tagged (55+75)
  - 12 SMEs: Each cluster tagged by 3 SMEs
    - SMEs assigned a number for anonymity
  - 52 Clusters were tagged 3 times
  - 39 Clusters were tagged 6 times

Cluster Analysis		
55		**75**
39	*Identical*	*39*
16	13 x 2 2 x 3 1 x 4	36

Clusters 55-24 & 75-31
Identical Subdomains
• fdic.gov • fdicconnect.gov • fdicig.gov • fdicoig.gov • fdicseguro.gov • myfdicinsurance.gov • egrpra.gov

Tag Analysis

- How topically related are the tags?
- Two researchers independently assigned "relatedness category" (RC)
  - 1 = little or no relation
  - 2 = somewhat related
  - 3 = strongly related

Cluster 55-19
2 Subdomains
• federalregister.gov • fedreg.gov

Cluster 55-19	SME 40	SME 32	SME 42
RC 3	• federal regulations • administrative law	• federal regulations	• federal regulations

Category 1: Very Little or No Relatedness

- Cluster 55-16

SME 35	SME 31	SME 39
• Geography • Government purchasing • Industrial safety • Intelligence service. • Small business.	• NONE	• federal regulations

- acqnet.gov
- acquisition.gov
- arnet.gov
- chemsafety.gov
- cia.gov
- csb.gov
- dia.mil
- dmso.mil
- fbo.gov
- fedbizopps.gov
- fedteds.gov
- lsc.gov
- myfloridahouse.gov
- nro.gov
- nrojr.gov
- odci.gov
- osdbu.gov
- stennis.gov
- tda.gov
- truman.gov
- uscapitolpolice.gov
- ustda.gov

Category 2: Somewhat Related

- Cluster 75-37

SME 3	SME 37	SME 38
• Hazardous substances -- Accidents -- Investigation -- United States. • Legal aid -- United States. • United States. Capitol Police	• public service education • Public Service Leadership	• chemical safety • Public Service Leadership

- chemsafety.gov
- csb.gov
- lsc.gov
- myfloridahouse.gov
- stennis.gov
- truman.gov
- uscapitolpolice.gov

Category 3: Strongly Related

- Cluster 55-18

SME 38	SME 39	SME 42
• Banks and Banking -- United States • Federal Deposit Insurance Corporation • financial industry regulation	• Banks and Banking -- United States • Federal Deposit Insurance Corporation • Bank Fraud -- United States	• Banks and Banking -- United States

- egrpra.gov
- fdic.gov
- fdicconnect.gov
- fdicig.gov
- fdicoig.gov
- fdicseguro.gov
- myfdicinsurance.gov

Findings: Tag Analysis

- Results: Relatedness Categories (*N* = 130)
  - 1 = little or no relation (*n* = 27; 21%)
  - 2 = somewhat related (*n* = 24; 18%)
  - 3 = strongly related (*n* = 79; 61%)
- Cluster Analysis successfully identified topically related subdomains in 61% of clusters

Clusters	1	2	3
130	21%	18%	61%
75-Set	21%	17%	61%
55-Set	20%	20%	60%

39 Identical Clusters

Analysis of Cluster Tagging Exercise

Cluster Analysis			Tagging Exercise
55		**75**	**130 clusters**
39	*Identical*	*39*	*Tagged 6 times*
16	13 x 2 2 x 3 1 x 4	36	Tagged 3 times

13 clusters: Six SMEs
21 clusters: Five SMEs
5 clusters: Four SMEs:

Same SME tagged the cluster twice

Clusters 55-46 & 75-63
3 Subdomains
• usccr.gov • fmcs.gov • adr.gov

Consistency Analysis: 39 Clusters

Clusters 55-46 & 75-63
3 Subdomains
• usccr.gov • fmcs.gov • adr.gov

Cluster 55-46	SME 40	SME 32	SME 31
RC 3	• mediation • dispute resolution	• mediation	• Mediation and conciliation, Industrial
Cluster 75-63	SME 35	SME 32	SME 31
RC 2	• Dispute resolution (Law) • Collective bargaining -- United States • Civil rights • Human rights	• mediation • dispute resolution	• Mediation and conciliation, Industrial

Consistency Analysis: 39 Clusters

- Each cluster pair had two RC values
  - 74% of RC values were the same (*n* = 29)
  - 26% of RC values were different (*n* = 10)
- Reevaluated 10 clusters
  - 7 Clusters: RC values of 2 and 3
  - 3 Clusters: RC values of 1 and 3
- Results
  - 7 Clusters: All were recoded as 3
  - 3 Clusters: Recoded as 1, 2, or 3
    1. Recoded as 1: 55-44/75-59
    2. Recoded as 2: 55-43/75-58
    3. Recoded as 3: 55-40/75-53

Findings: 39 Clusters

- Suggests that more taggers allow for more consistent assessments of subdomain relatedness within a cluster
  - More than 3 taggers might be better!
- Tags from 4-6 SMEs impacted RC assessments
  - Fewer in RC 2
  - More in RC 3

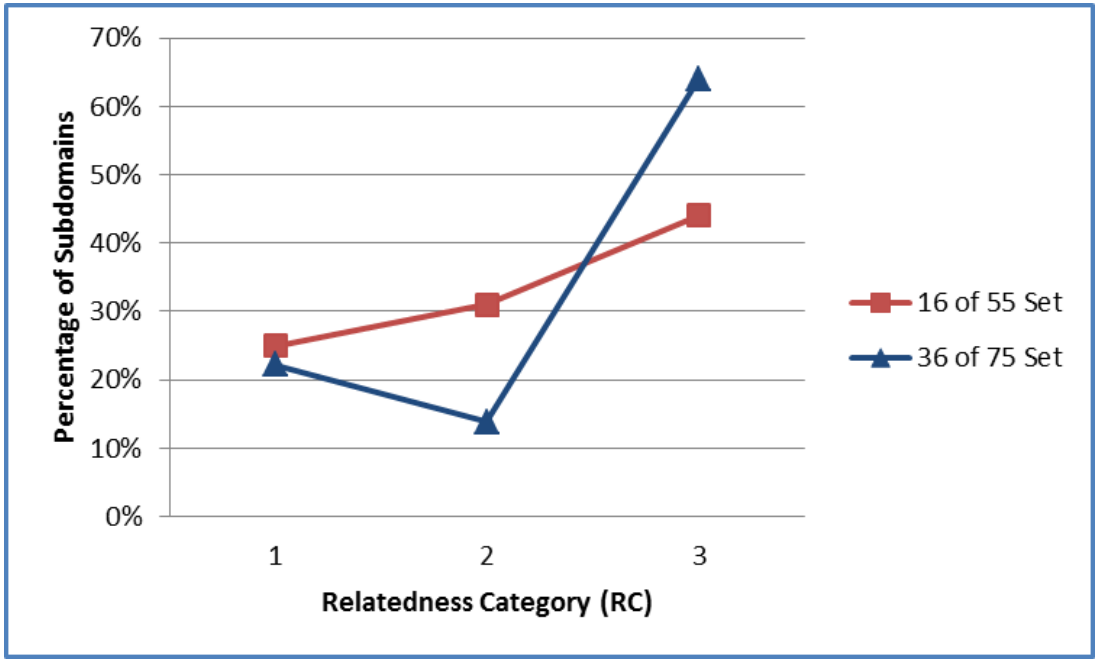
Cluster Set	RC 1	RC 2	RC 3
130	21%	18%	61%
39	18%	10%	72%

Impact of Increasing the Number of Clusters

Impact of Increasing Number of Clusters

55-16	1	3	2	
55-22	1	3	1	
55-10	1	2	1	
55-54	1	2	1	
55-38	2	3	3	1
55-21	2	3	3	
55-33	2	3	2	
55-41	2	3	2	
55-7	2	3	2	1
55-26	3	3	3	3
55-5	3	3	3	
55-8	3	3	3	
55-13	3	3	3	
55-47	3	3	3	
55-6	3	3	1	
55-49	3	3	1	

From 16 Clusters to 36 Clusters



Impact of Increasing Number of Clusters

Clusters	# Subdomains	RC 1	RC 2	RC 3
Combined	130	21%	18%	61%
Identical	39	18%	10%	72%
55-Set	16	25%	31%	44%
75-Set	36	22%	14%	64%

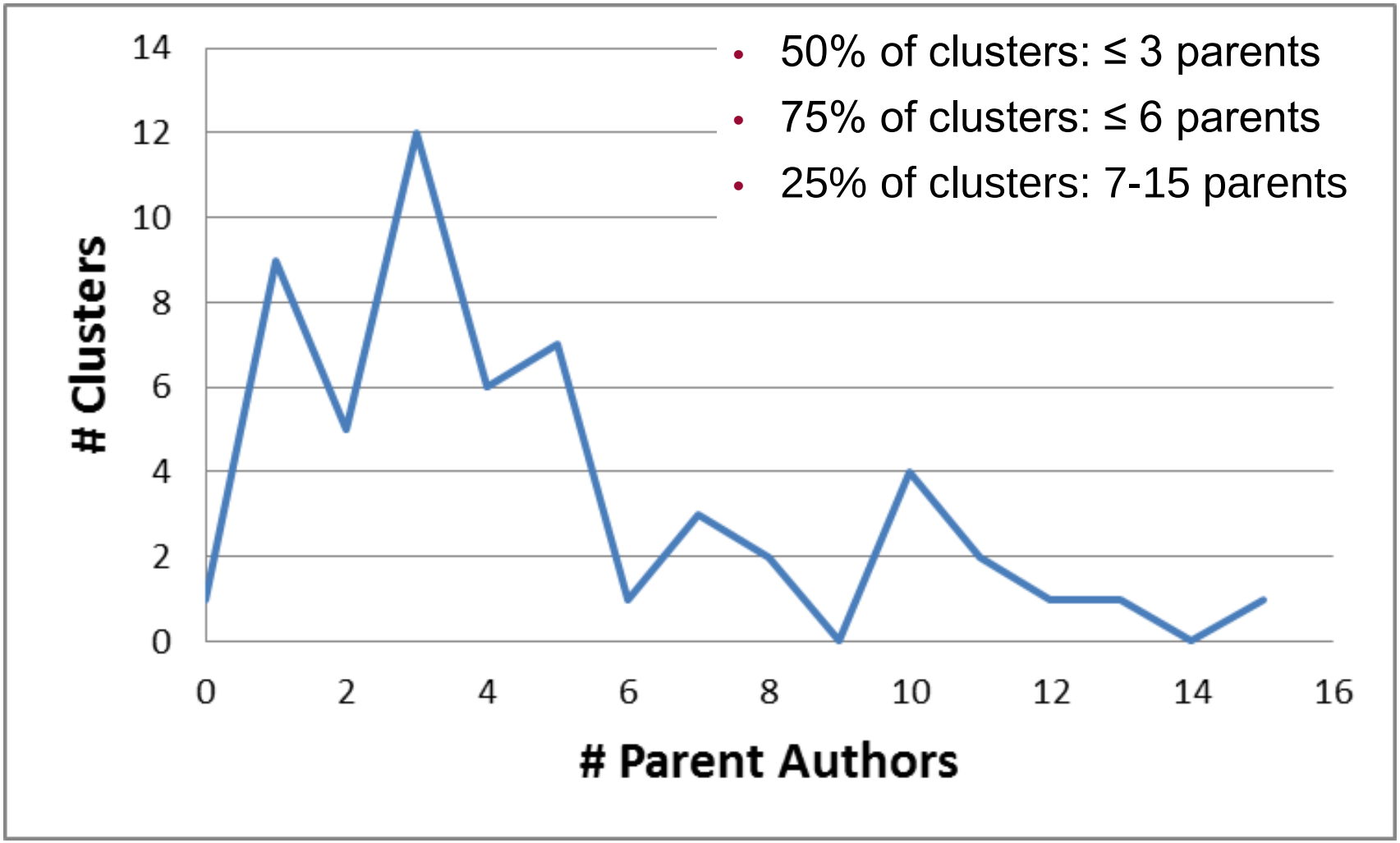
- Clusters that remained intact (i.e., 39 identical clusters in both 55-set and 75-set) had the highest percentage of topically related subdomains
  - RC 3: 72% v. 61%
- Clusters that separated into smaller clusters (16 into 36) had a higher percentage of topically related subdomains after the break-up
  - RC 3: 64% v. 44%

Overall Findings

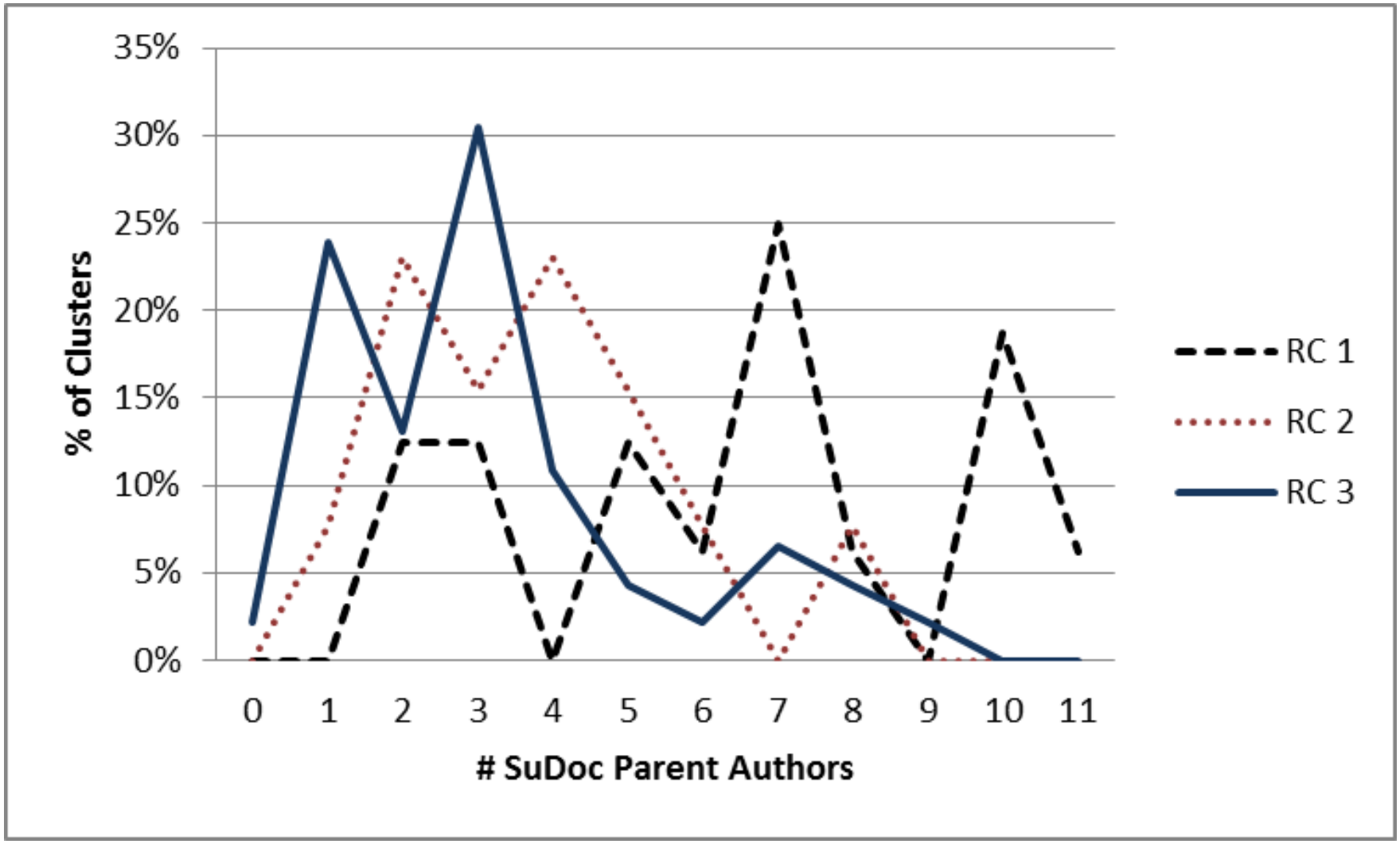
Clusters, SuDocs, & RCs

RC	1	2	3
CLUSTERS (*N* = 75)	16	13	46
# Subdomains			
average	15	12	16
range	3-48	3-30	2-53
# SuDoc Authors			
average	8	6	6
range	2-16	2-14	0-15
# SuDoc Parents			
average	6	4	3
range	2-11	1-8	0-9

SuDoc Classification of Subdomains: 55 Clusters



Findings: Tagging Exercise



METRICS

Metrics: Methods

- Focus group discussion with project's SMEs
  - Identify criteria used for acquisition of materials from Web archives
- Survey of FDLP Libraries
  - Purpose: Assess libraries' interests and capabilities in accessing v. acquiring content from Web archives
  - Participants: 414 libraries in the Federal Depository Library Program
- Review of current statistics and measurement

Metrics: Focus Group Findings

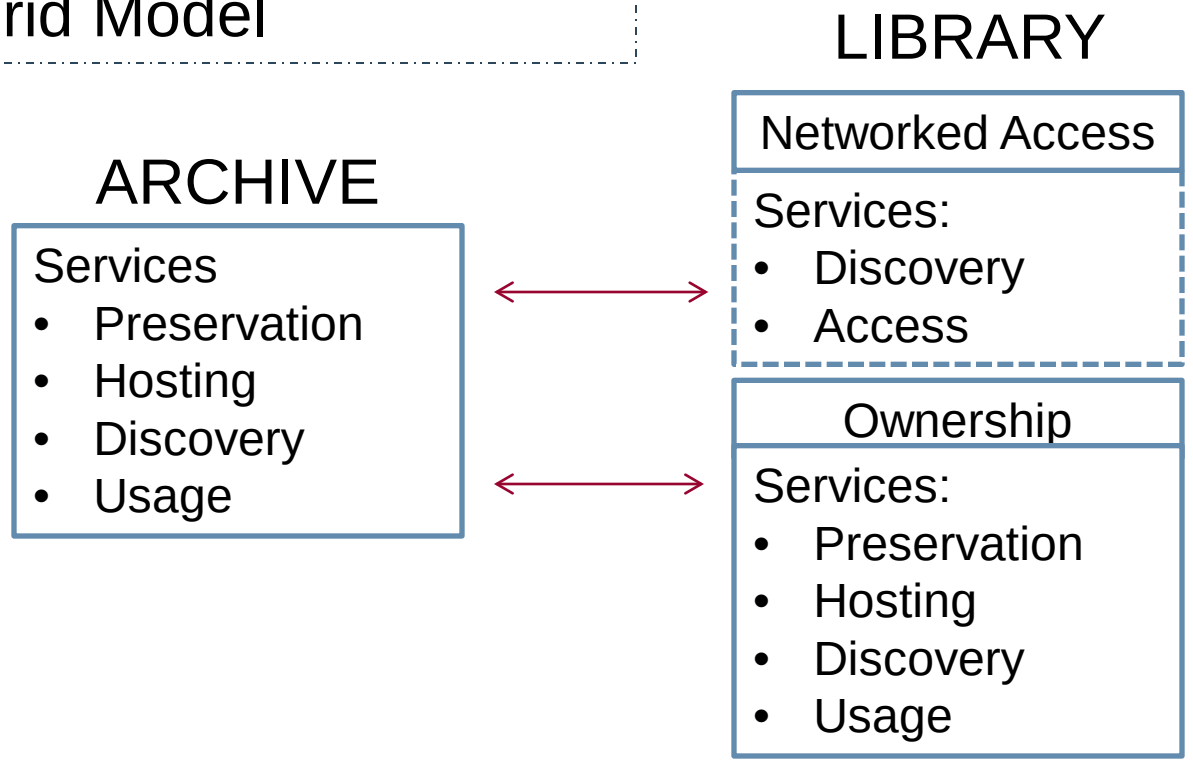
- More libraries interested in networked access to an archive v. purchasing and hosting locally
- Current metrics for networked electronic resources are best informants for Web archive content
  - Critical importance of standards-compliant usage data
- Authorities - Standards
  - ARL; ACRL; NCES/IPEDS
  - COUNTER: Codes of Practice
    - Counting Online Usage of Networked Electronic Resources
    - SUSHI: ANSI/NISO Z39.93-2007
      - Standardized Usage Harvesting Initiative

Metrics: Focus Group Findings

- Content description informs selection decisions
  - Topical areas covered
  - Unique or exclusive content available
  - Dates materials were harvested
- Metrics that drive acquisitions
  - Retention: Cost per use
  - Selection: Usage data (when available)
- Categories of statistics and measurements
  - Scope (How much; how many)
  - Expenditures (Cost)
  - Usage (Counts)
  - Quality (Outcomes; Impacts; Value)

Metrics: Web Archive Service Models

1. Networked Access Model
2. Ownership Model
3. Hybrid Model



Metrics: Proposed Statistics
SCOPE

- For a Web archive:
  - Size (in gigabytes, terabytes, etc.)
  - Number of discrete collections
- For each collection within a Web archive:
  - Size (in gigabytes, terabytes, etc.)
  - Number of objects by type:

Text	109,498,363	Dataset	908,339
Image	29,140,868	Video	318,498
Text-like	11,234,522	Audio	198,349
Computer file	3,472,193		

Metrics: Proposed Statistics
USAGE

- For each collection within a Web archive:
  - Number of sessions
    - Total number
    - Number federated or automated
  - Number of searches (queries)
    - Total number of searches run
    - Number federated or automated

CLOSING

EOTCD Project Accomplishments

- Selection of Materials in Web Archives
  - PROBLEM: Foreknowledge of a resource's URL is often required
  - PROBLEM: The absence of descriptive metadata or classification schemes thwarts discovery & access
  - RESULT: A solid basis for further investigation of cluster analysis, particularly when combined with SME involvement, as an organizational mechanism to enhance resource discovery

EOTCD Project Accomplishments

- Metrics for Materials in Web Archives
  - PROBLEM: Acquisition & retention decisions require standard metrics which are not available
  - RESULT: Unique contribution to the metrics needed from the librarian's perspective, particularly in the areas of content description, scope, and usage

What's Next

- Full-text search
  - How do we integrate what we've learned?
  - What other improvements to Web archive search can we make?
- Using the Web graph
  - How do we leverage the graph for identifying content?
- Describing the collection
  - How can we engage faculty with our Web archives?
- Identifying change
  - How is the .gov Web changing over time?

Thanks!

Kathleen Murray
kathleen.murray@unt.edu

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

- Reports & Presentations