Curation of the End-of-Term Web Archive

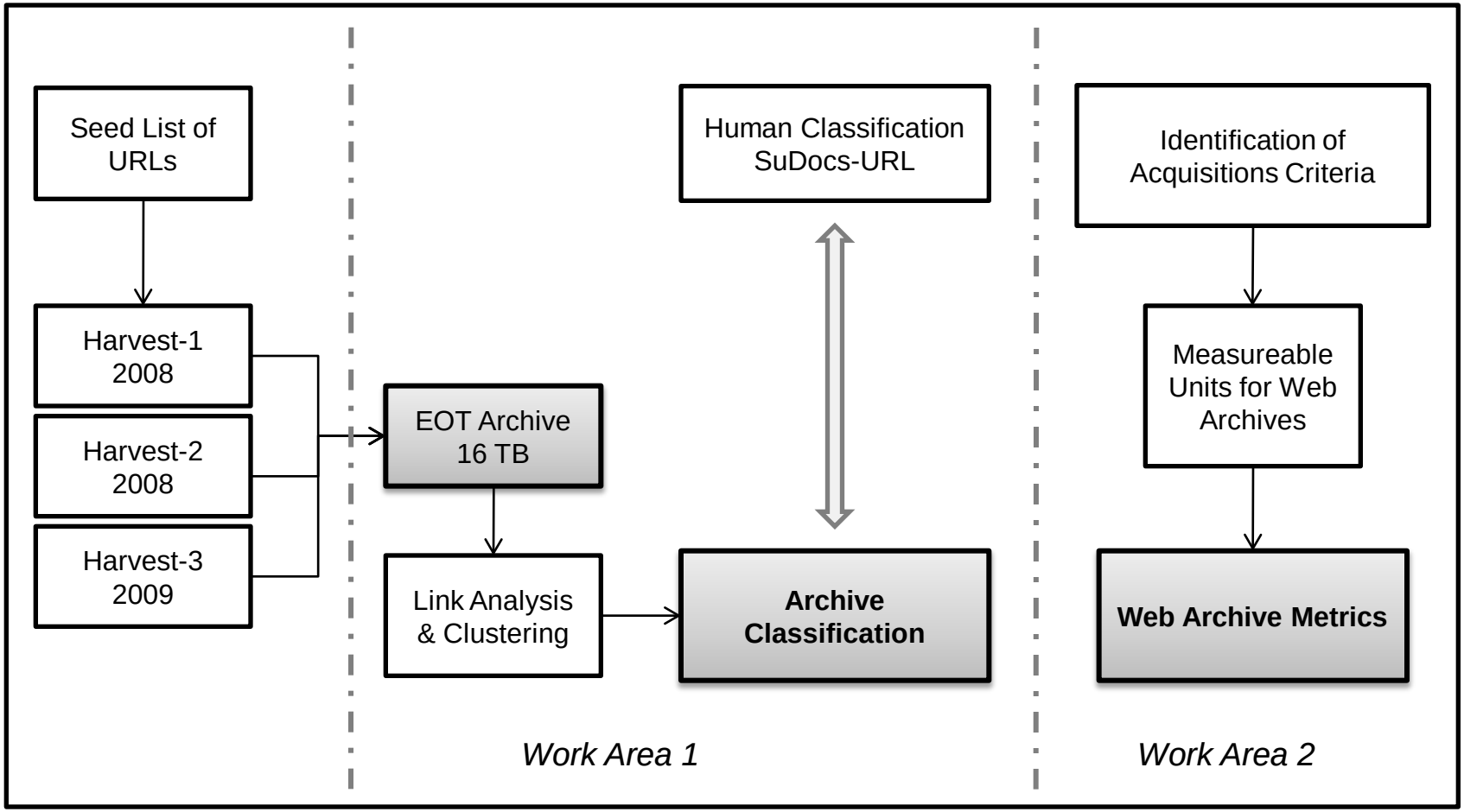
Kathleen Murray – University of North Texas Libraries

Advisory Board Meeting – November 4, 2011

Topics

- Background
  - EOTCD Project
- Findings
  - Archive Classification
    - Human classification v. cluster analysis
    - Cluster tagging
  - Metrics for Web Archives
- Discussion: What's Next?

Background: EOTCD Work Areas



ARCHIVE CLASSIFICATION

Classification: Size Challenges

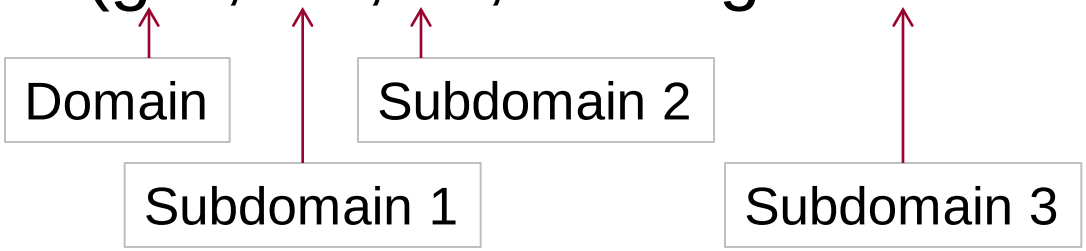
Largest Domains	# URLs	# Unique Subdomains
gov	137,847,822	14,339
com	7,809,711	57,873
org	5,108,645	29,798
mil	3,555,425	1,677
edu	3,552,509	13,856

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

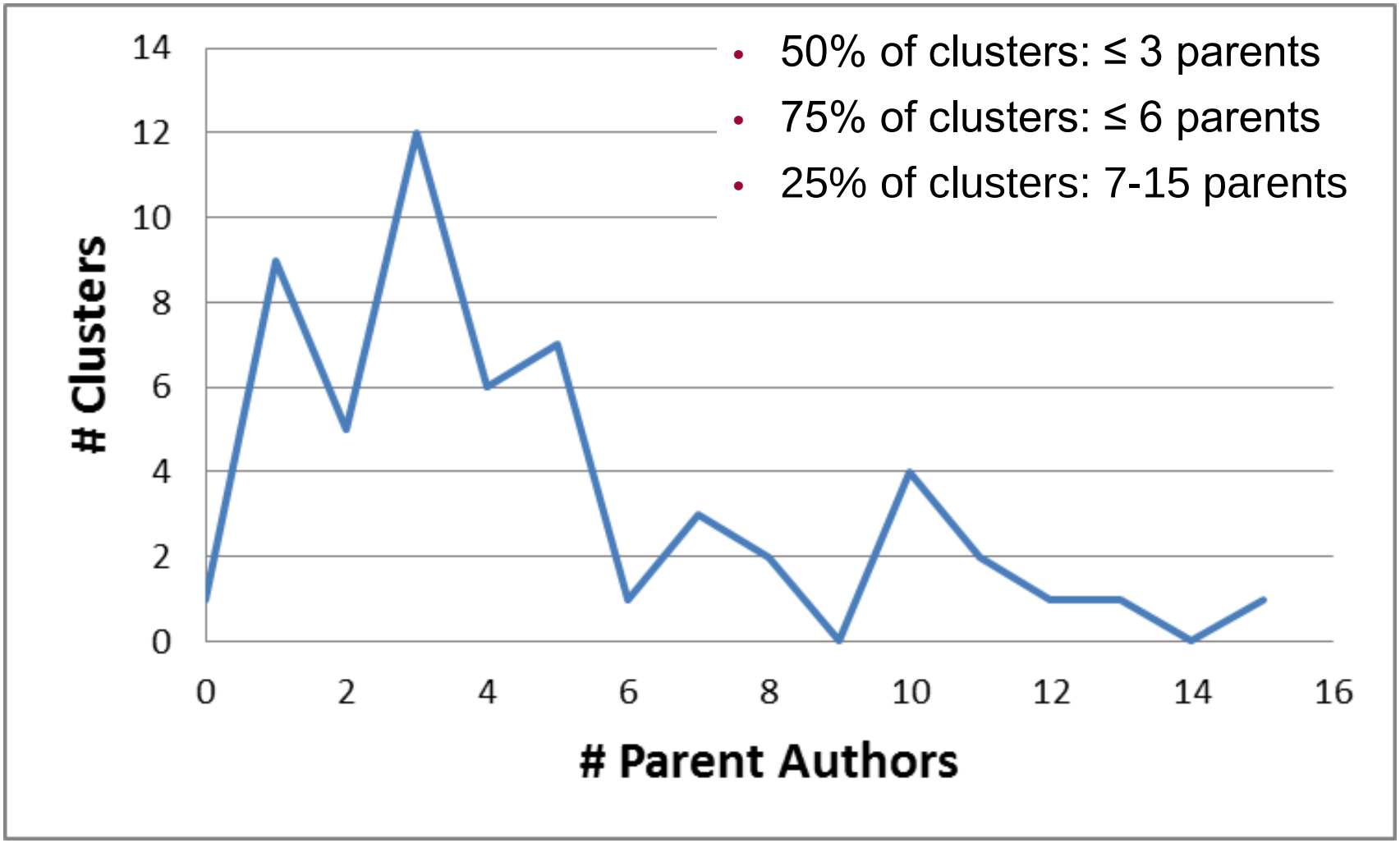
- SuDocs Classification System
- 10 SMEs classified 1,151 Web sites corresponding to the 1,151 subdomains
  - Each site classified by 2 SMEs
  - 70% agreement (*n* = 808); 30% disagreement (*n* = 343)
- 3 arbitrators classified 343 Web sites
- Final result:
  - Assigned SuDocs authors to 1,040 subdomains
    - 1,111 authors (1,040 + 71 multiply authored sites)
  - Unable to assign SuDoc authors to:
    - 60 sites: within scope of federal government
    - 51 sites: out of federal government scope

Cluster Analysis

- Utilized the Web graph
- A number of cluster analysis algorithms were explored
  - Best result: Agglomerative Hierarchical Clustering
- Set limit on number of clusters to identify
  - First analysis: Set of 55 clusters
  - Second analysis: Set of 75 clusters

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

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 government agency authors
  - In the case of multiple authors, there were often 1-2 dominant authors
  - Implication for search results:
    - May be feasible to suggest related sites within the Archive 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
  - 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

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%

Impact of Increasing Number of Taggers

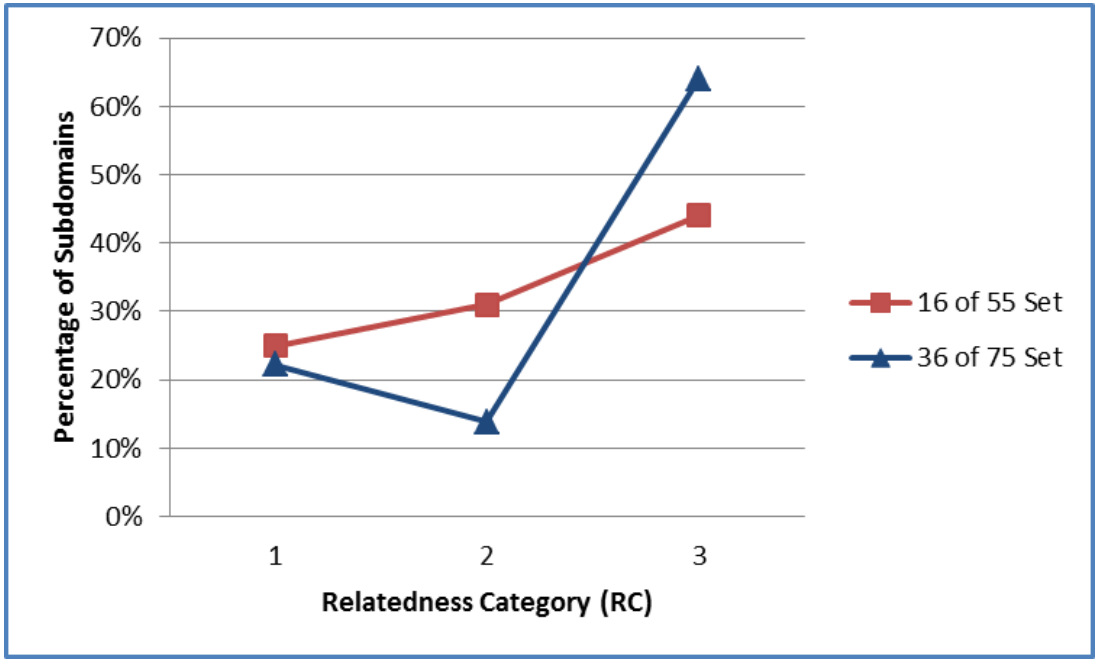
Cluster Set	RC 1	RC 2	RC 3
130	21%	18%	61%
39	18%	10%	72%

- 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 30

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



Conclusions

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%

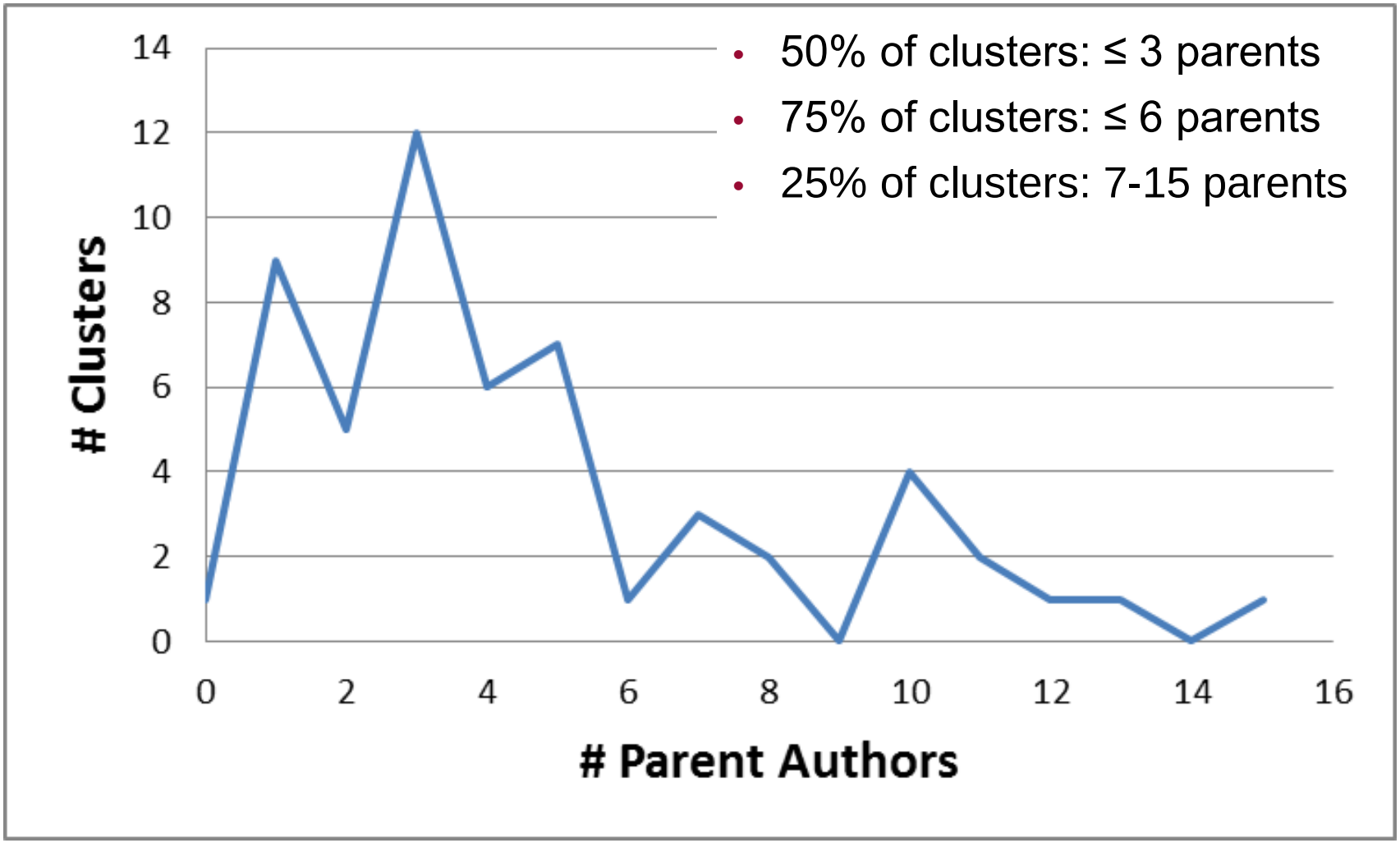
- Clusters that remained intact (i.e., 39 identical clusters in both 55-set and 75-set) had the highest percentage of topically related subdomains
  - RC 3: 72% v. 61%
- Clusters that separated into smaller clusters (16 into 36) had a higher percentage of topically related subdomains after the break-up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

Overall Findings

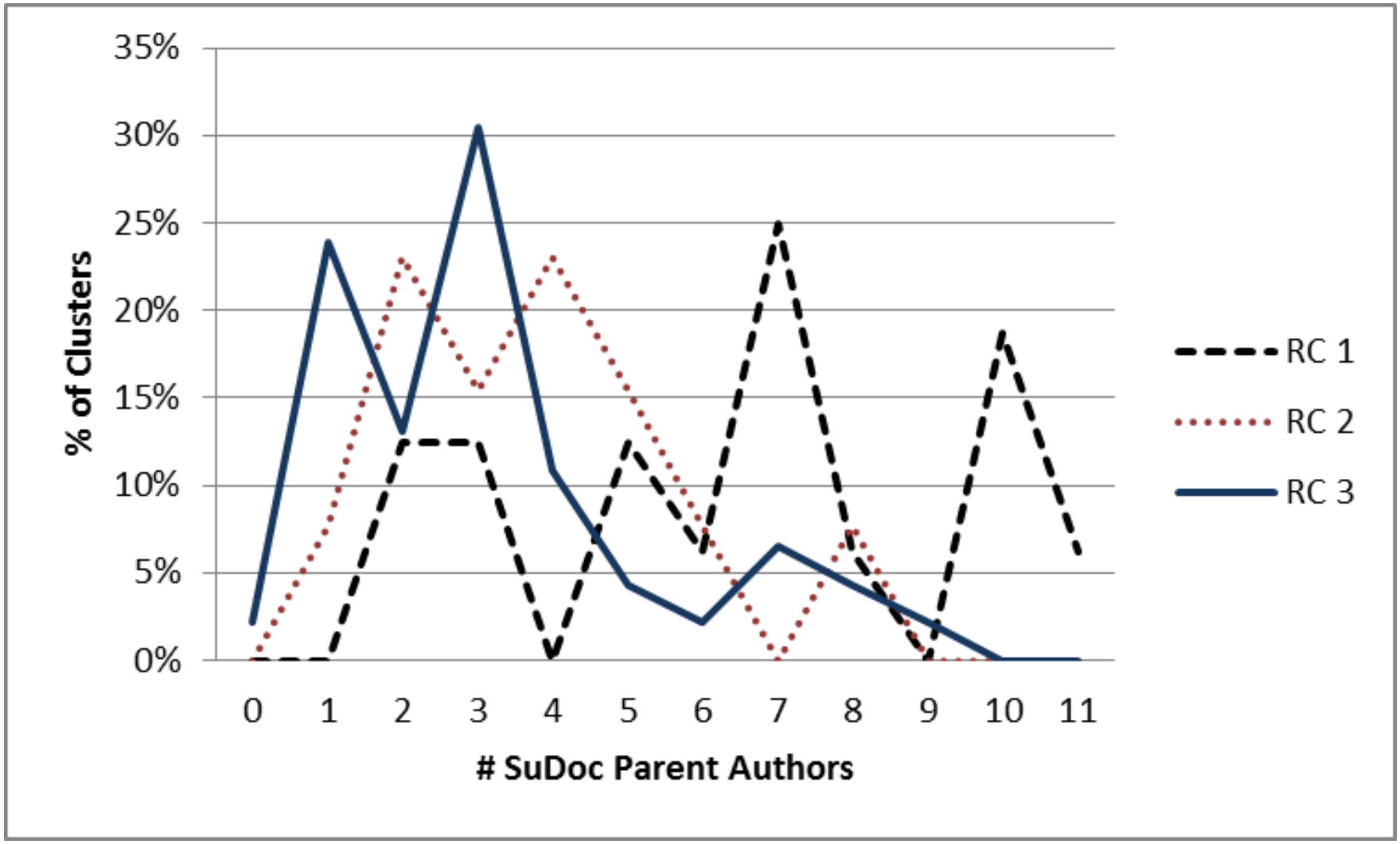
Clusters, SuDocs, & Relatedness (RC)

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 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: 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
Document-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

- EOT Archive Classification
  - PROBLEM:
    - The absence of descriptive metadata or classification schemes thwarts discovery & access
    - Foreknowledge of a resource's URL is often required
  - OBJECTIVE: Classify materials in accord with the SuDocs Classification Numbering System
    - To enable librarians to utilize existing selection practices to identify materials in the EOT Archive
  - RESULT: A solid basis for further investigation of cluster analysis to enhance resource discovery
    - Particularly when combined with SME involvement

EOTCD Project Accomplishments

- Metrics for Materials in Web Archives
  - PROBLEM: Acquisition & retention decisions require standard metrics which are not available
  - OBJECTIVE: Identify a set of metrics for materials in Web archives
    - To enable characterization of materials in Web archives in units of measurement more familiar to libraries and their administrations
  - 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?

END