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

SME Meeting – October 16, 2011 – Washington DC

UNT Libraries THE POWER OF IDEAS STARTS HERE IMLS Award LG-06-09-0174-09

Agenda

1:00 PM	Working Lunch – Project Update
1:30 PM	Cluster Tagging
2:15 PM	Break
2:30 PM	Focus Group Discussion
3:45 PM	Closing Remarks

Topics

- Background
- Cluster Tagging
  - Examples of Relatedness Sub-Categories
  - Results
  - 39 Identical Clusters
  - Impact of Increasing the Number of Clusters
- Overall Findings
  - SuDoc Classification & Tagging
- What's Next
- Focus Group Discussion
- Closing Remarks

Background

Classification: Challenges

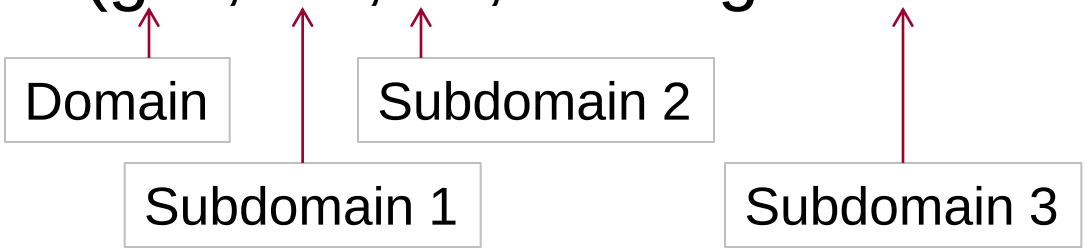
Largest Domains	# URLs	# Unique Subdomains
gov	137,847,822	14,339
com	7,809,711	57,873
org	5,108,645	29,798
mil	3,555,425	1,677
edu	3,552,509	13,856

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

Link Analysis: Web Graph

- 1,151 subdomains
  - Multiple URLs per subdomain
  - Example: Library of Congress (LOC) - 44 URLs
    - SURTs format:
      - http://(gov,loc,)
      - http://(gov,loc,catalog,)
      - http://(gov,loc,webarchive,)
- Link extraction: 62,452 links inter-relating HTML files
  - Includes outlinks and inlinks for each URL
- Each pair of linked subdomains assigned a weight
  - Reflecting the number of actual links between the URLs in each source/target subdomain pair

Cluster Analysis

- A number of cluster analysis algorithms were explored
  - Best result to date: Agglomerative Hierarchical Clustering
- Set limit on the 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

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)
  - Unable to classify 111 subdomains (in/out of scope)

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?
- 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

Sub-categories of Relatedness
Selected Examples

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 1.1

- All clusters tagged “NONE”
- Cluster 75-70

SME 31	SME 40	SME 43
• NONE	• NONE	• NONE

- achp.gov
- africanburialground.gov
- cendi.gov
- dnfsb.gov
- exim.gov
- fcsic.gov
- hoopa-nsn.gov
- iawg.gov
- imls.gov
- nlrb.gov
- recdata.gov
- rfets.gov
- sdp.gov
- ustr.gov

Category 1.2

- Two SMEs tagged “NONE”; one with keywords
- Cluster 75-29

35	38	39
• NONE	• labor • Social security -- United States • U.S. Consumer Product Safety Commission.	• NONE

- atvsafety.gov
- cpsc.gov
- dea.gov
- directoasucuenta.gov
- fbiic.gov
- gao.gov
- godirect.gov
- medpac.gov
- mspb.gov
- nmb.gov
- segurosocial.gov
- socialsecurity.gov
- ssa.gov
- wdol.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 2.1

- Two SMEs tagged the cluster with related keywords; one SME tagged with “NONE”
- Cluster 55-35

SME 34	SME 35	SME 32
• aviation research • polar research • scientific research	• National Science Foundation (U.S.) • Polar regions Research • Research	• NONE

- arctic.gov
- faa.gov
- faasafety.gov
- gsadvantage.gov
- itrd.gov
- microbeproject.gov
- nano.gov
- nitrd.gov
- nsf.gov
- research.gov
- usap.gov
- us-ipy.gov

Category 3: Strongly Related

- Cluster 55-18

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

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

Category 3.1

- Strong relationship; one SME added many additional tags
- Cluster 55-11

- accessmanagement.gov
- boosterseat.gov
- bts.gov
- cflhd.gov
- cmts.gov
- dot.gov
- fightgridlocknow.gov
- fmip.gov
- italladdsup.gov
- mrcog-nm.gov
- nhtsa.gov
- ntdprogram.gov
- plainlanguage.gov
- protectyourmove.gov
- safercar.gov
- safercars.gov
- safertruck.gov
- safertrucks.gov
- tfhrc.gov
- topnet.gov
- transportation.gov

Category 3.1 - Cluster 55-11 con't.

SME 33	SME 32	SME 38
• car pools • child car seats • Child restraint systems in automobiles -- United States • child safety • Emergency prepardness • Roads -- United States • Shipping -- United States • Telecommuting • terrorist threat • Traffic congestion--Government policy--United States. • Transportation -- United States -- Statistics • transportation information • Trucks -- Safety measures	• Transportation	• Transportation • United States. Department of Transportation.

Category 3.2

- Core of strongly related tags with one SME adding moderate amount of additional tags
- Cluster 55-28

SME 33	SME 38	SME 37
• Law -- Databases. • legal research • Libraries-- United States • Library of Congress • United States -- History • United States -- Politics and government	• Libraries-- United States	• Library of Congress

- americaslibrary.gov
- americasstory.gov
- americastory.gov
- crs.gov
- glin.gov
- loc.gov

Category 3.3

- One SME's tags were a superset of the other two
- Cluster 75-53

SME 34	SME 42	SME 36
• Economic Data • Economic development • International trade	• Economic Data • Foreign trade -- United States • Foreign trade -- United States – Statistics	• Foreign trade -- United States -- Statistics

- economy.gov
- eurotradeonline.gov
- oecdonline.gov
- stat-usa.gov
- usatradeonline.gov
- useconomy.gov

Results of the Tagging Exercise

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

- Total of 91 unique clusters tagged
  - 39 Identical clusters that were tagged by 6 SMEs
  - 52 clusters that were tagged by 3 SMEs

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

39 Identical Clusters: Consistency Analysis

- Intra-tagger reliability: 26 Clusters
  - *21 Clusters*: 5 taggers
    - One SME tagged each cluster twice
  - *5 Clusters*: 4 taggers
    - Two SMEs tagged each cluster twice
- 31 cases of same SME tagging same cluster
  - Consistency measured on scale of 1-3
  - 97% consistency rate

Clusters 55-3 & 75-8
4 Subdomains
• arpa.gov • arpa.mil • darpa.mil • darpa.gov

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 assigned different RC values

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

Example: Different RC Values (3 and 1)

- Cluster 55-44
  - 37 Subdomains

- arts.gov
- californiadesert.gov
- cfa.gov
- dhra.mil
- dmg.gov
- dss.mil
- ecr.gov
- eklutna-nsn.gov
- espanol.gov
- faq.gov
- fca.gov
- fec.gov
- ferc.gov
- fireplan.gov
- firstgov.gov
- forestsandrangelands.gov
- gobiernousa.gov
- gov.gov
- government.gov
- govinfo.gov
- govtinfo.gov
- itds.gov
- listovirginia.gov
- mojavedata.gov
- ncix.gov
- nea.gov
- nonprofit.gov
- seniors.gov
- statelocal.gov
- udall.gov
- us.gov
- usa.gov
- usagov.gov
- usgov.gov
- usgovernment.gov
- usgovt.gov
- webcontent.gov

Example: Different RC Values (3 and 1)

Cluster 55-44	SME 34	SME 38	SME 42
RC 3	• Government publications -- United States. • general information search systems • Recreation areas -- United States • arts and humanities support	• U.S. Government information • Government publications -- United States	• U.S. Government information

Example: Different RC Values (3 and 1)

Cluster 75-59	SME 34	SME 43	SME 42
RC 1	• NONE	• NONE	• Environmentalism • public lands

Example: Different RC Values (3 and 1)

Cluster 55-44	SME 34	SME 38	SME 42
RC 3	• Government publications -- United States. • general information search systems • Recreation areas -- United States • arts and humanities support	• U.S. Government information • Government publications -- United States	• U.S. Government information
Cluster 75-59	SME 34	SME 43	SME 42
RC 1	• NONE	• NONE	• Environmentalism • public lands

What RC would you assign to this cluster?

Example: Different RC Values (3 and 1)

Cluster 55-44	SME 34	SME 38	SME 42
RC 3	• Government publications -- United States. • general information search systems • Recreation areas -- United States • arts and humanities support	• U.S. Government information • Government publications --United States	• U.S. Government information
Cluster 75-59	SME 34	SME 43	SME 42
RC 1	• NONE	• NONE	• Environmentalism • public lands

Poor intra-rater reliability indicates RC 1

Results of Reevaluation of 10 Clusters

- Each of the 10 clusters was initially assigned a different RC value
  - 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

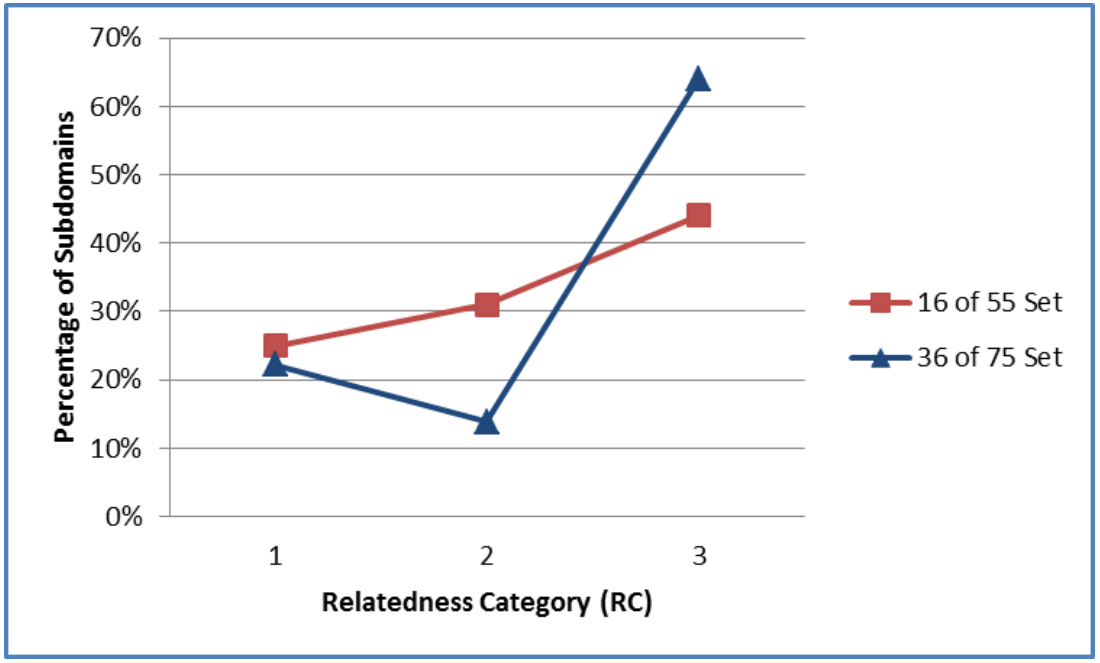
Clusters	# Subdomains	RC 1	RC 2	RC 3
Combined	130	21%	18%	61%
Identical	39	18%	10%	72%
55-Set	16	25%	31%	44%
75-Set	36	22%	14%	64%

- 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%

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

55-16	1	3	2
55-22	1	3	1
55-10	1	2	1
55-54	1	2	1

- 55-22 (RC 1); 28 Subdomains
  - 75-0 (RC 3); 14 Subdomains
  - 75-29 (RC 1); 14 Subdomains

Cluster 75-0		
SME 34	SME 38	SME 39
• People with disabilities • Discrimination in employment.	• People with disabilities	• People with disabilities • American disability act • department of justice • inspectors general

- abilityone.gov
- access-board.gov
- counterterrorismtraining.gov
- disabilities.gov
- fasab.gov
- fedsfeedfamilies.gov
- fmc.gov
- ignet.gov
- info.gov
- jwod.gov
- ncd.gov
- nigc.gov
- telework.gov
- uspsoig.gov

Impact of Increasing Number of Clusters

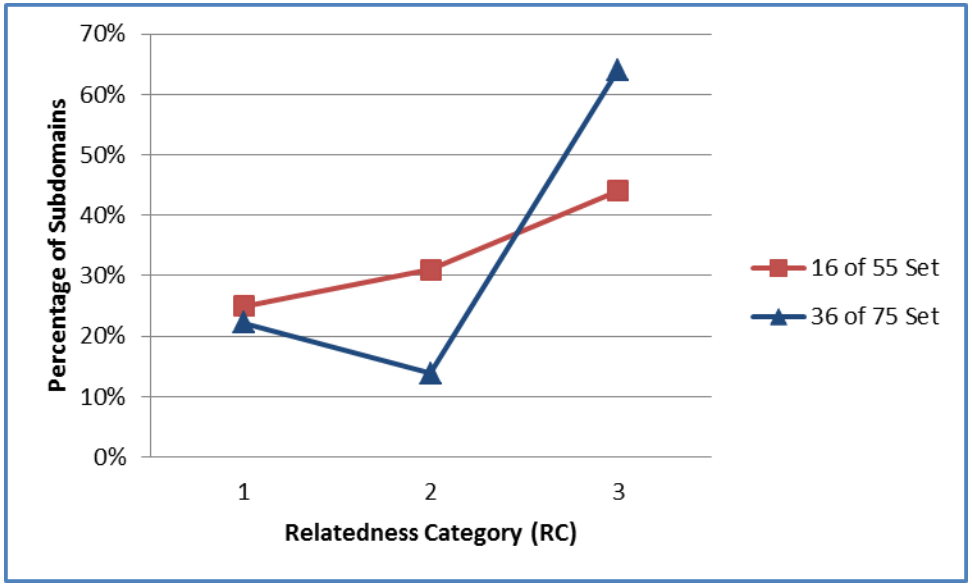
55-16	1	3	2
55-22	1	3	1
55-10	1	2	1
55-54	1	2	1

- 55-22 (RC 1); 28 Subdomains
  - 75-0 (RC 3); 14 Subdomains
  - 75-29 (RC 1); 14 Subdomains

Cluster 75-29		
SME 35	SME 38	SME 39
• NONE	• labor • Social security -- United States • U.S. Consumer Product Safety Commission.	• NONE

- atvsafety.gov
- cpsc.gov
- dea.gov
- directoasucuenta.gov
- fbiic.gov
- gao.gov
- godirect.gov
- medpac.gov
- mspb.gov
- nmb.gov
- segurosocial.gov
- socialsecurity.gov
- ssa.gov
- wdol.gov

Impact of Increasing Number of Clusters



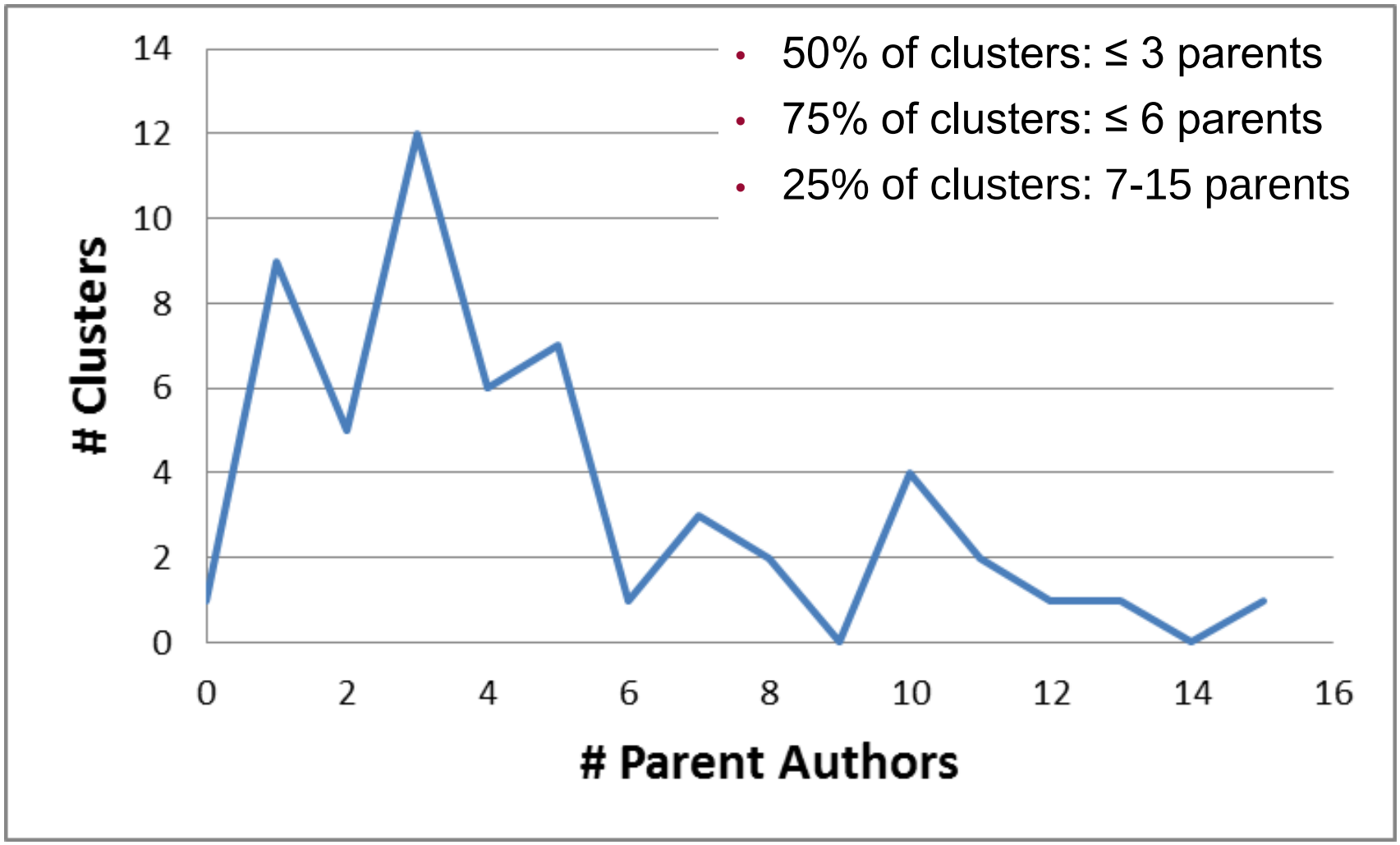
- Clusters that break into smaller clusters appear to identify:
  - Clusters whose subdomains are more topically related (RC 2 → RC 3)
  - Clusters whose subdomains are topically unrelated (RC 1)

Overall Findings

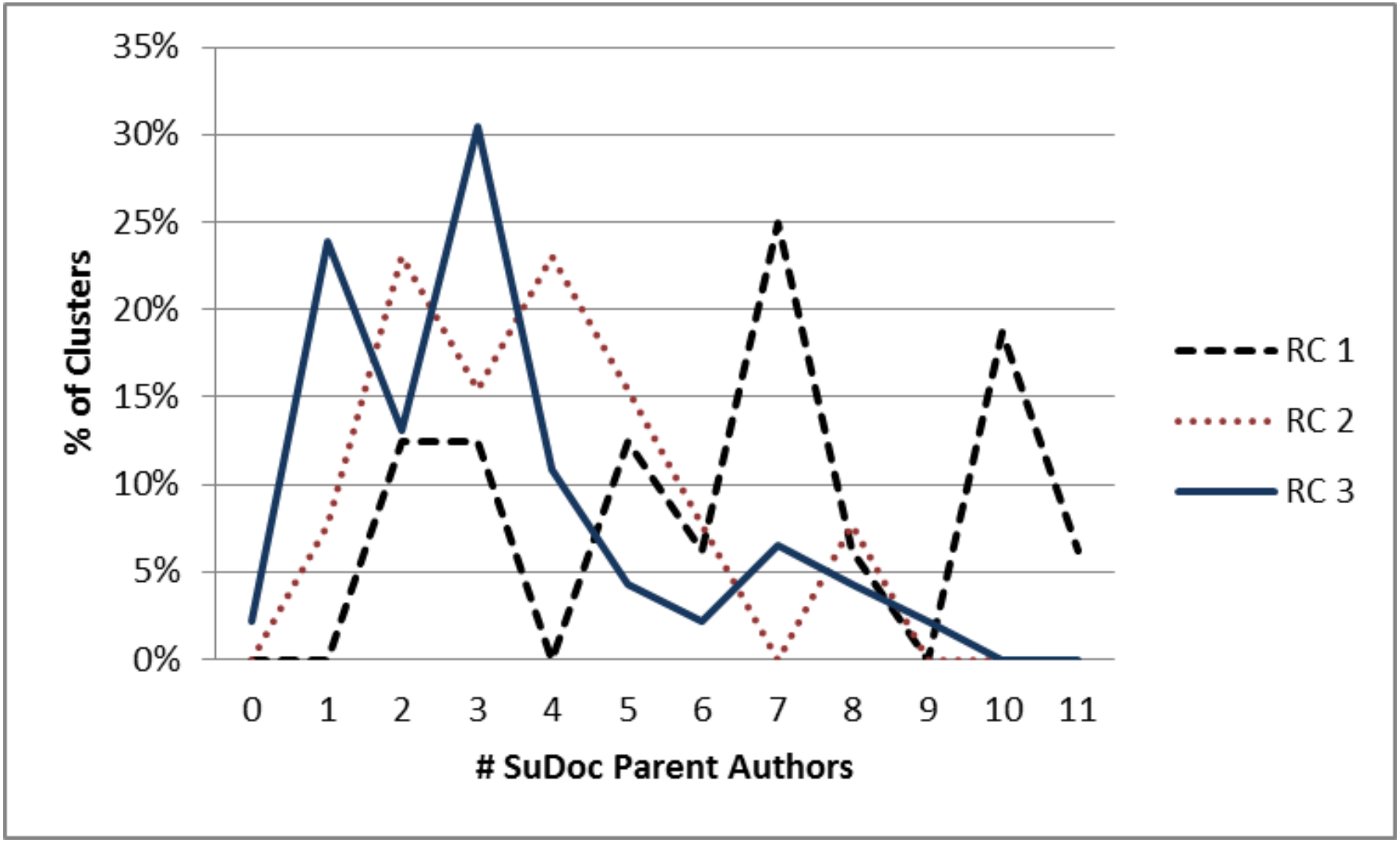
Clusters, SuDocs, & Relationship Categories

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

Subdomain Classification: 55 Clusters



Findings: Tagging Exercise



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 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?

Focus Group Discussion

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

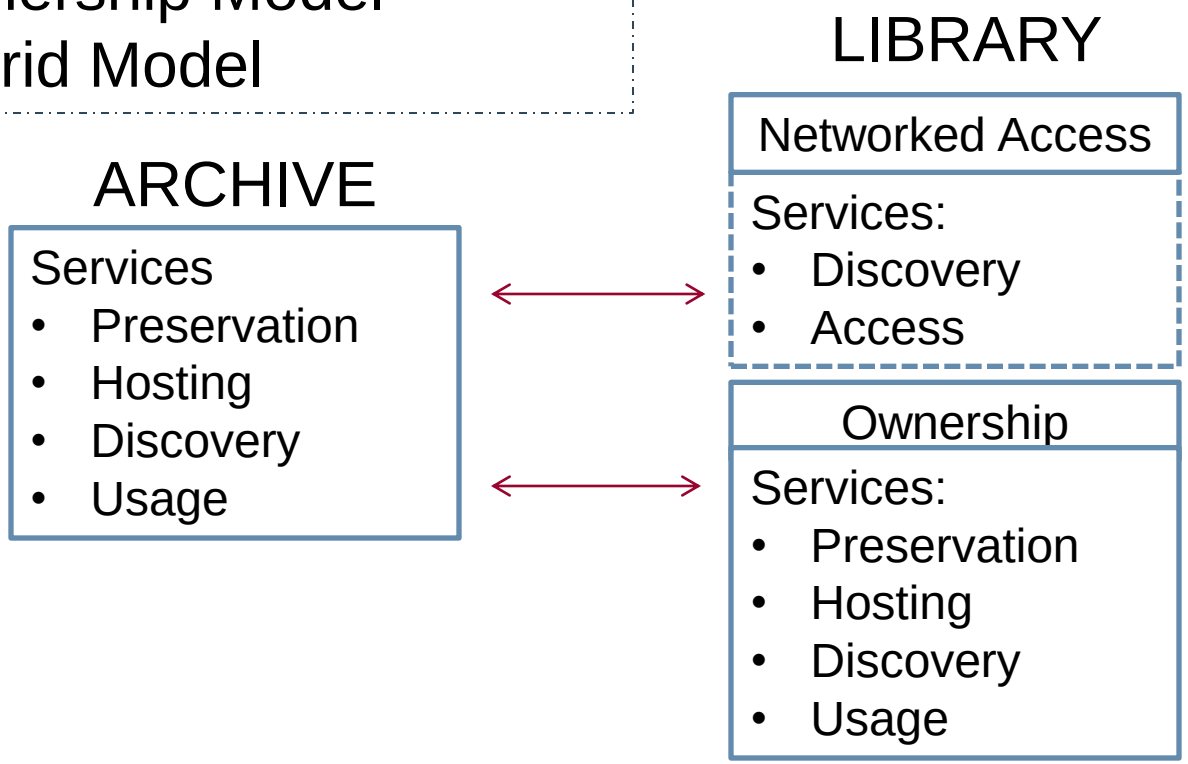
- More libraries interested in networked access to an archive v. purchasing and hosting locally
- Current metrics for networked electronic resources are best informants for Web archive content
  - Critical importance of standards compliant usage data
- Authorities - Standards
  - ARL; ACRL; NCES/IPEDS
  - 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)

A. Metrics: Web Archive Service Models

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



Are these models adequate for libraries?

B. Metrics: Content Description

- Content description informs selection decisions
  - Topical areas covered
  - Unique or exclusive content available
  - Dates materials were harvested

- Do these meet your needs for resource selection?

- What additional information about a collection would be helpful to you?

C. Metrics: Proposed Structure
COSTS

- Provision of both free and fee-based services
- Example: A tiered cost structure from service providers:
  - Free basic discovery and access services
  - Fee-based options and services:
    - usage reports
    - hosting

- In general, do you think libraries are willing to pay for services beyond basic discovery and access?
- What if these services are from archives hosting web-published resources harvested from federal government agencies?

D. Metrics: Proposed Statistics

SCOPE – *Materials Held by Library*

- 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:

EXAMPLE: EOT ARCHIVE			
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		

Will these address statistical reporting needs?

E. 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

- Will these Counter-compliant usage statistics satisfy libraries' requirements?
- What additional usage data do you think would be useful?

F. Metrics: Possibility
QUALITY

- UK Serials Group has been investigating a journal usage factor as a measure of the quality and value of online journals.

- How do you think usage data from a Web archive could be used as a measure of quality?

- What dimensions of quality do you think will be important to libraries in regard to Web archives, their collections, and materials?

G. Metrics: Usage Reports

- Emulate the COUNTER usage reports for databases and journals. As such they would include:
  - Sessions by Month by Collection
  - Searches by Month by Collection
  - Searches and Sessions by Year by Collection
  - Searches and Sessions by Year by Archive
- As appropriate, these reports could be done for consortia as well as individual institution.

- Will these Counter-compliant reports satisfy libraries' requirements?

H. Resource Discovery

- Looking ahead to selecting resources for your collection from a Web archive such as the End-of-Term Archive:

- What are some of the pros and cons of discovering resources using:
  - URLs
  - SuDoc stems
  - Subject tags (keywords)
- If only one of these options was possible, which would you prefer?

I. Web Archives

- Looking back on your experience since our first project meeting in Buffalo in April 2009:

- How has your understanding of Web archives changed over the last two years?

- Is your understanding more clear or more muddled?

Please take a few moments to complete a brief questionnaire!

Closing Remarks

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

Closing

- Project Website http://research.library.unt.edu/eotcd
  - Reports and presentations available now
- UNT Digital Library http://digital.library.unt.edu/
  - Reports preserved for future access
- Expense Reports
  - Please submit to Cathy Hartman as soon as possible

Thanks very much for your participation!