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

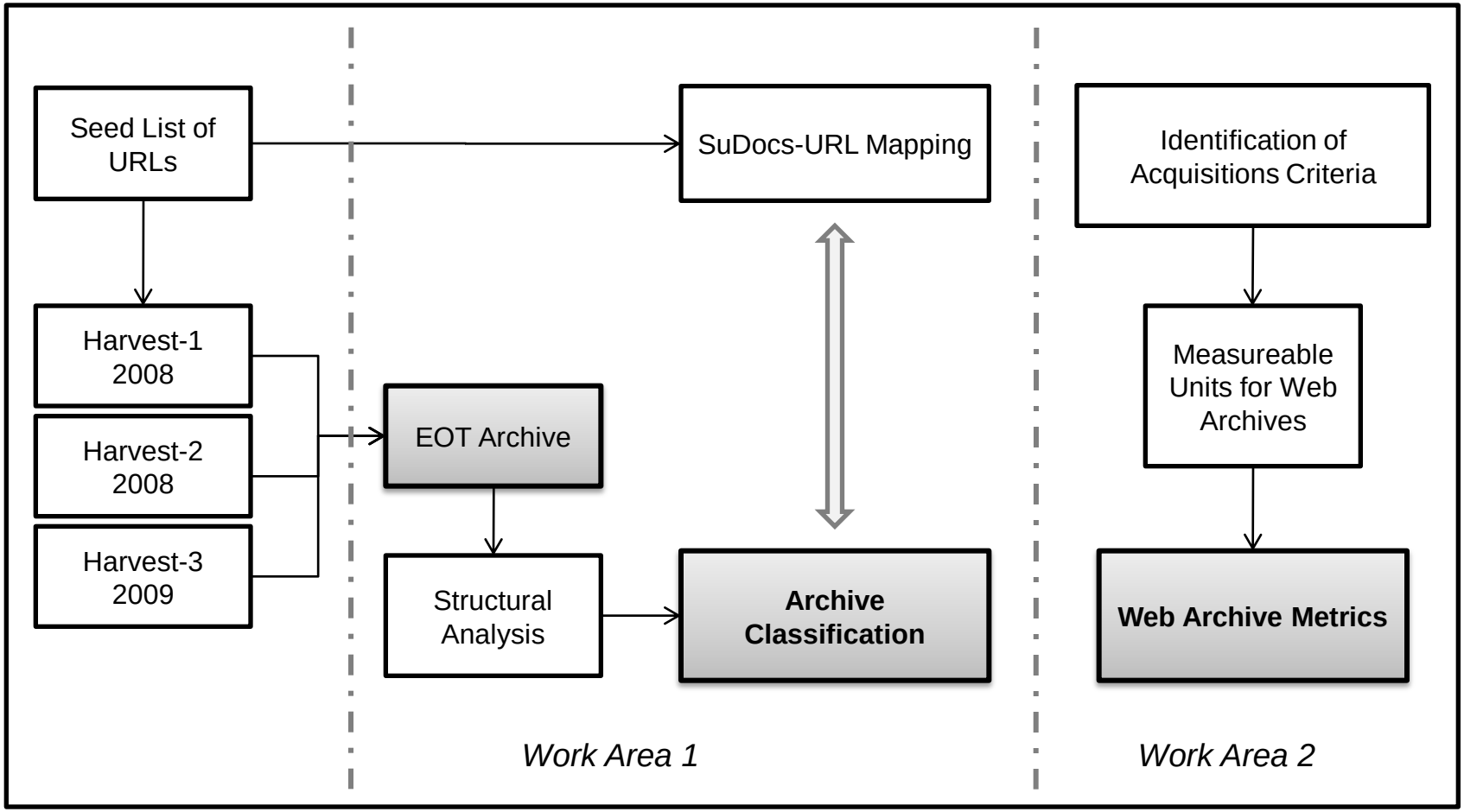
SME Meeting – April 3, 2011 – San Antonio

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

Agenda

11:30 AM	Working Lunch – Project Status
12:00 PM	Archive Classification Results
1:00 PM	Break
1:15 PM	Archive Classification Results
1:45 PM	Link Analysis Results: Clusters
2:30 PM	Closing Remarks
2:45 PM	End

Project Status



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

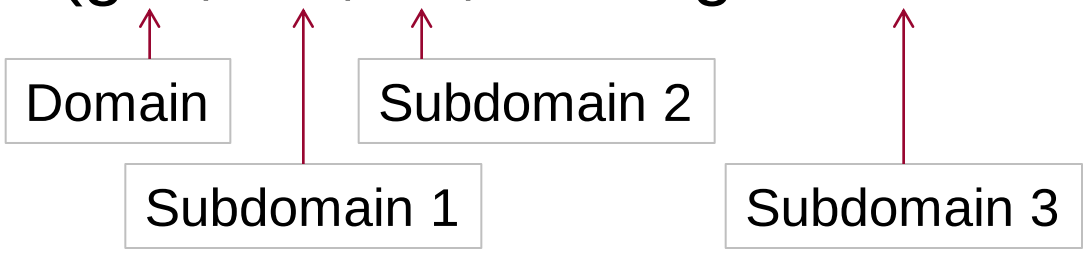
Reduced Unique Subdomains to 16,015

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

Initial Classification

- Sample of 1,151 URLs in End-of-Term Archive
  - Unique subdomains within the .mil & .gov domains
- 10 SMEs classified Websites
  - Each classified by 2 SMEs: 230/person
  - Average time spent: 10 hours 45 minutes (*n* = 6)
- Results
  - 70% agreement (*n* = 808)
    - Unable to classify:
      - 18 - in scope
      - 36 - out of scope
  - 30% disagreement (*n* = 343)

Categories of Disagreements

Category	#	%	% Sample
Additional author(s)	110	32%	10%
Classified v. in scope	68	20%	6%
No agreement	66	19%	6%
Parent v. subordinate	56	16%	5%
Classified v. out of scope	36	10%	3%
In scope v. out of scope	7	2%	1%
	343	100%	30%

Feedback

- Overall
  - Classification tool was easy to use
  - Exercise was fun and educational: Discovered agencies
- SuDocs Classification System
  - Overall, it worked well to classify Websites
  - Lacks sufficient granularity for subordinate offices and agencies
    - Departments of Energy & Defense
    - Native American sites
  - Forced to classify at high level
- Classification Challenges
  - Major challenge: Determining primary author
  - Server hosting page (GSA in particular) not the author of content

Multiple Authors

Disagreements			Sample
Category	#	%	%
Additional author(s)	110	32%	10%

- Strategies to determine the primary author
  - URL; host server; “contact us”; first or largest agency logo
  - One person reported guidance in the *Catalog of US Government Publications* (CGP) useful; another knew of no established hierarchy for Web-published materials

Multiple Authors

- Unable to identify primary author
  - If 2-3 agencies, included them all
  - If > 3 agencies:
    - Classified as “In scope – unable to classify”
    - Wanted a “working group” classification; existing interagency classes did not suffice
- Suggestions:
  - Establish a multi-agency stem (MA)
  - Establish series designations for digital object types:
    - Databases
    - Audio recordings; video recordings
    - Blogs

Arbitrator Classification

- 343 Websites in End-of-Term Archive
  - Subdomains SMEs classified differently
- 3 arbitrators
  - Each classified 114 Websites
  - Evaluated SME classifications, including notes
- Results
  - Assigned SuDocs authors to 286 Websites
  - Unable to classify 57 Websites
    - In scope: 42
    - Out of scope: 15

Arbitration Results

Arbitrator Action	#	%
Selected one SME classification	281	82%
Picked new classification	43	13%
Selected one SME classification & added authors	10	3%
Picked one author from multiple SME authors	7	2%
Picked one author from multiple SME authors	7	2%

Multiple Authors

- SME Agreements
  - Three authors
    - watermonitor.gov
    - nationalresourcedirectory.gov
  - Two authors
    - time.gov
    - vitm.gov
    - telework.gov
- Arbitrator Decisions
  - Five authors: tradeagreements.gov
  - Four authors: nehrp.gov

Classification Examples

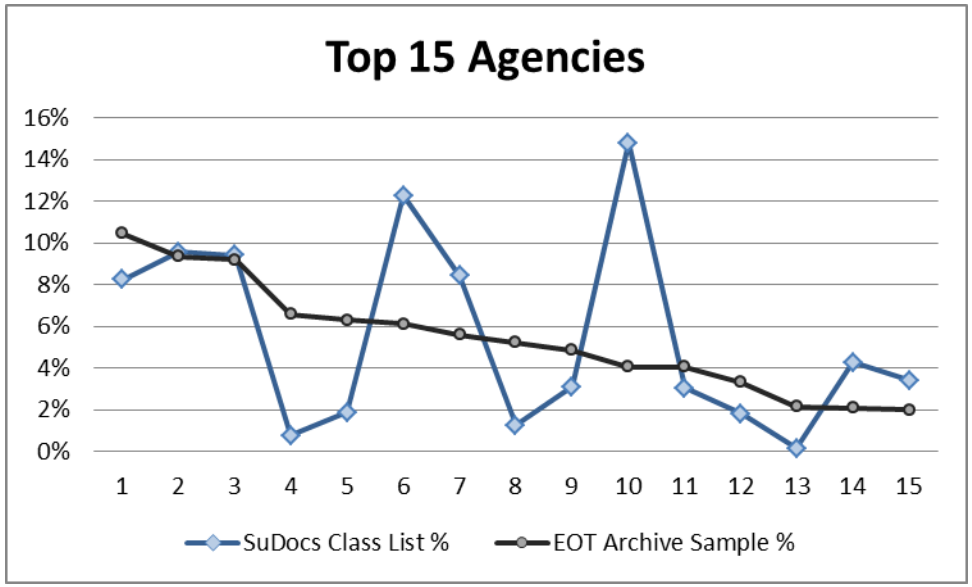
- Additional author(s)
  - firescience.gov
  - californiadesert.gov
- Classified v. in scope
  - acquisition.gov
  - execsec.gov
- No agreement
  - identitytheft.gov
  - manufacturing.gov
  - africanburialground.gov
- Parent v. subordinate
  - airnow.gov
  - health.gov
- Classified v. out of scope
  - dra.gov
  - housedemocrats.gov
- In scope v. out of scope
  - cdatribe-nsn.gov
  - mitigationcommission.gov

BREAK

Final Classification

- Multiple Authors: 56 Websites (71 authors)
  - Five: 1 Websites
  - Four: 3 Websites
  - Three: 11 Websites
  - Two: 41 Websites
- Unable to classify: 111 Websites
  - In scope: 60 Websites
  - Out of scope: 51 Websites
- Final count:
  - 1,040 Websites assigned SuDocs stems
  - 1,111 authors (1,040 + 71)

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

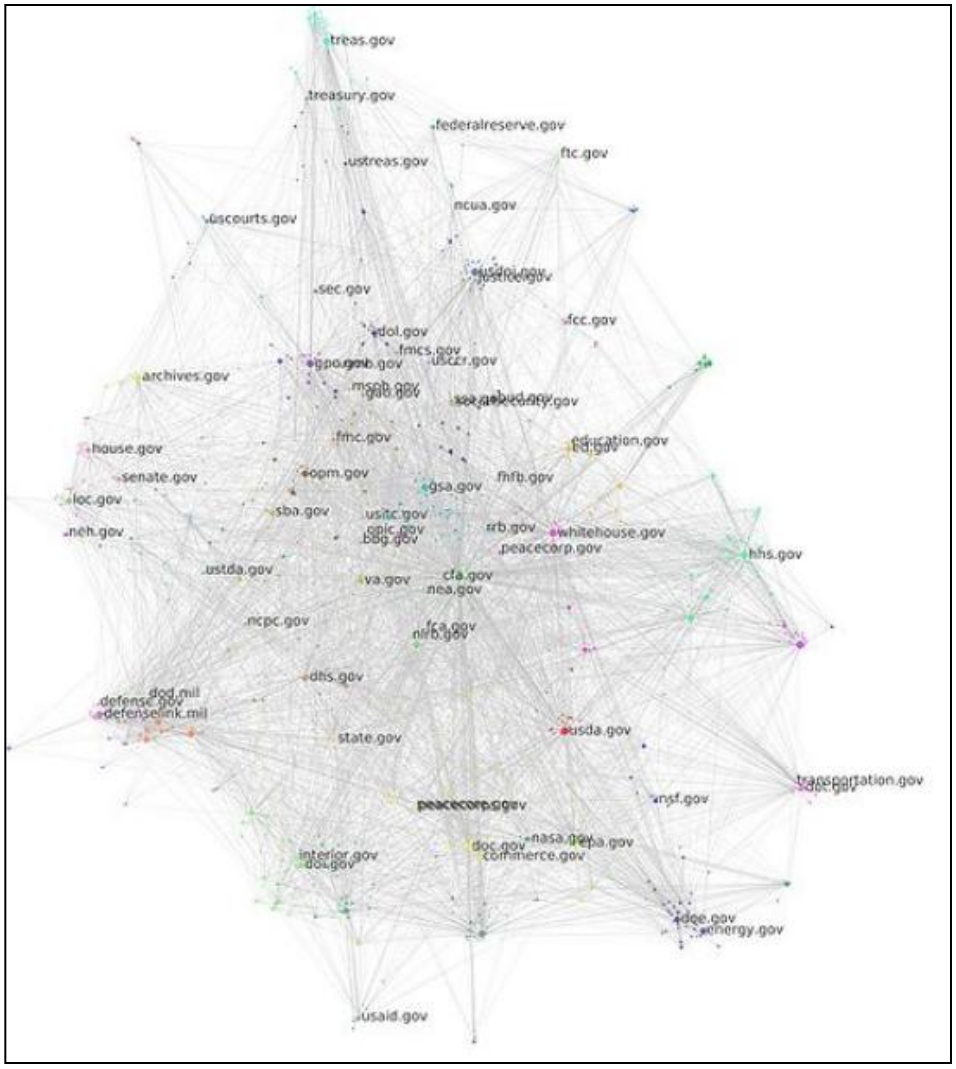
Link Analysis & Clustering

- Clustering methods

$$\mathrm{NGD}(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

  - LinLog Clustering
  - Linlog Coordinates With Agglomerative Hierarchical Clustering
  - Normalized Google Distance (NGD)
  - Strongest Outlinks and Majority Inlinks
  - Web Communities
- Optimal Clusters (*at this point*)
  - Linlog Coordinates With Agglomerative Hierarchical Clustering
    - 55 Clusters
    - 75 Clusters

Visualization of Clusters



Linlog Coordinates with Agglomerative Hierarchical Clustering

- Limited to 55 clusters
- Force-directed visualization with Protovis
  - Parent agency subdomains identified
  - Colors correspond to agency

Cluster Evaluation

- SuDocs stems assigned to clusters: 55 & 75 clusters
- SME evaluation of clusters
  - Three people will evaluate each cluster (*N* = 130)
    - Identify subject terms to describe content
    - Identify misfits
- Exercise: Subject Tag Tool
  - Enter *subject tags*
  - Timeframe: Summer 2011
- Outcome
  - Feedback to refine the cluster analysis
  - Folksonomy to describe web-published content

Closing

- Web Archive Metrics
  - Item Selection Profiles for SME Libraries
  - Identifying sites within EOT Archive consistent w/ profiles
- Project Website
  - http://research.library.unt.edu/eotcd
    - Reports & updates
    - Work in progress
- Expense Reports
- Next SME Meetings
  - October 2011: Washington DC

Thanks very much for your participation!