PRESIDENT | VICE PRESIDENT | FIRST LADY | MRS. CHENEY | NEWS

Your Government | History & Tours | Kids | E-mail | En Español

THE WHITE HOUSE
PRESIDENT GEORGE W. BUSH

Search

Podcasts RSS Feeds

September 15, 2008

Last updated 6:16 p.m. (EDT)

IN FOCUS

Afghanistan
Africa
Budget Management
Defense
Economy
Education
Energy
Environment
Global Diplomacy
Health Care
Homeland Security
Immigration
International Trade
Iraq
Judicial Nominations
Middle East
National Security
Veterans

more issues

LATEST NEWS

+ more photos

White House Photo by Chris Greenberg

President George W. Bush and Mrs. Laura Bush, joined by President John Agyekum Kufuor and Mrs. Theresa Kufuor of Ghana, acknowledge the crowd Monday, Sept. 15, 2008, following the South Lawn Arrival Ceremony for President Kufuor and Mrs. Kufuor of Ghana at the White House. White House photo by Chris Greenberg

FEATURES

Ask the White House

Discuss Small Businesses and Health Savings Accounts (HSAs)
Join Sandy K. Baruah, Acting Administrator, U.S. Small Business Administration, on Thurday, September 18 at 12:00 PM to discuss small businesses and Health Savings Accounts (HSAs)

Submit a Question

Official State Visit

President and Mrs. Bush are pleased to welcome the President of the Republic of Ghana and Mrs. Kufuor for a State Visit on September 15, 2008. The United States and Ghana enjoy warm relations and a shared commitment to promote peace and prosperity in Africa and throughout the world.

Classifying the End of Term Web Archive

Kathleen Murray & Cathy Hartman

University of North Texas, UNT Libraries, 1155 Union Circle #305190, Denton, TX 76203-5017

E-mail ZIP GET UPDATES

the BRIEFING ROOM | ISSUES | the ADMINISTRATION | ABOUT the WHITE HOUSE | our GOVERNMENT | CONTACT us

THE ADMINISTRATION • PRESIDENT BARACK OBAMA

SEARCH

PRESIDENT BARACK OBAMA

Barack H. Obama is the 44th President of the United States.

His story is the American story — values from the heartland, a middle-class upbringing in a strong family, hard work and education as the means of getting ahead, and the conviction that a life so blessed should be lived in service to others.

With a father from Kenya and a mother from Kansas, President Obama was born in Hawaii on August 4, 1961. He was raised with help from his grandfather, who served in Patton's army, and his grandmother, who worked her way up from the secretarial pool to middle management at a bank.

After working his way through college with the help of scholarships and student loans, President Obama moved to Chicago, where he worked with a group of churches to help rebuild communities devastated by the closure of local steel plants.

He went on to attend law school, where he became the first African—American president of the *Harvard Law Review*. Upon graduation, he returned to Chicago to help lead a voter registration drive, teach constitutional law at the University of Chicago, and remain active in his community.

President Obama's years of public service are based around his unwavering belief in the ability to unite people around a politics of purpose. In the Illinois State Senate, he passed the first major ethics reform in 25 years, cut taxes for working families, and expanded health care for children and their parents. As a United States Senator, he reached across the aisle to pass groundbreaking lobbying reform, lock up the world's most dangerous weapons, and bring transparency to government by putting federal spending online.

Download Low-res (200KB)
Download Hi-res (1MB)

THE ADMINISTRATION

President Barack Obama
Vice President Joe Biden
First Lady Michelle Obama
Dr. Jill Biden
The Cabinet
White House Staff
Executive Office of the President

STAY CONNECTED

Facebook
YouTube
Twitter
Vimeo
Flickr
iTunes
MySpace

A NEW ERA of RESPONSIBILITY
FY 2010 BUDGET
MORE INFORMATION

YOUR MONEY at WORK
RECOVERY.GOV
MORE INFORMATION

Archive Background

- Captured US government's Web presence:
  - September 2008 - November 2009
  - Transition between George W. Bush & Barack Obama
  - 16 TB
  - 160 million URLs (files/documents)

Domains	# URLs	Subdomains
gov	137,847,822	14,339
mil	3,555,425	1,677

Problem

- Absence of descriptive metadata & Website classification thwarts discovery & access
- Standard file format (WARC - ISO 28500)
  - Specifies formats for storage, management, & exchange of data objects
  - Not designed for user access
- Access requires knowledge of a resource's URL (Wayback interface)

Objective

To enable government information librarians to utilize existing selection practices to identify materials in the EOT Archive

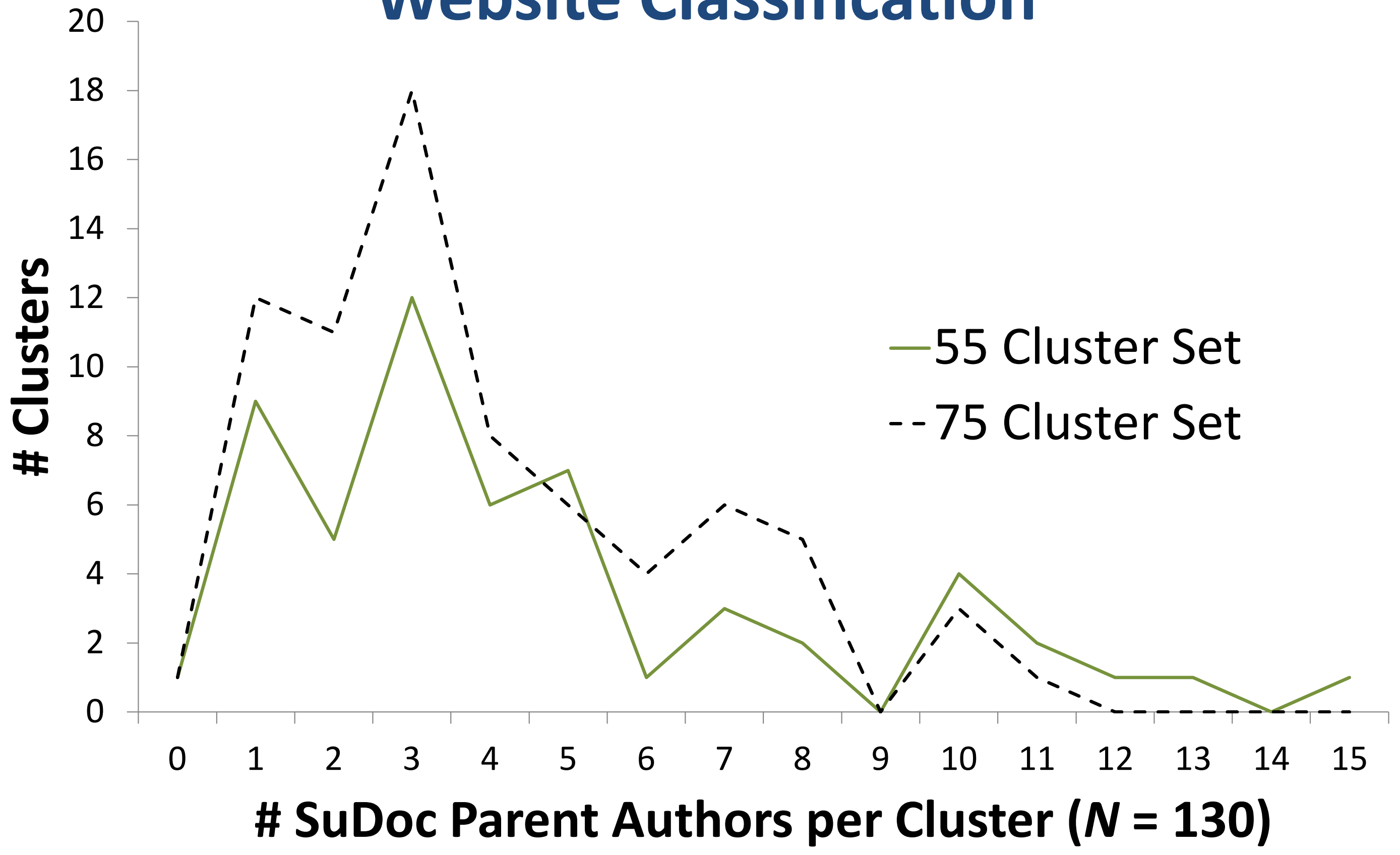
Methods

- Cluster Identification (1,151 subdomains)
  - Linlog Coordinates with Agglomerative Hierarchical Clustering
  - Two sets of related Websites (55-set & 75-set)
- Website Classification (1,151 subdomains)
  - Government publication classification scheme: Superintendent of Documents (SuDocs) Classification Numbering System
  - 10 Subject Matter Experts (SMEs) & 3 arbitrators
- Cluster Tagging
  - 12 SMEs assigned subject terms to clusters

Website Clusters

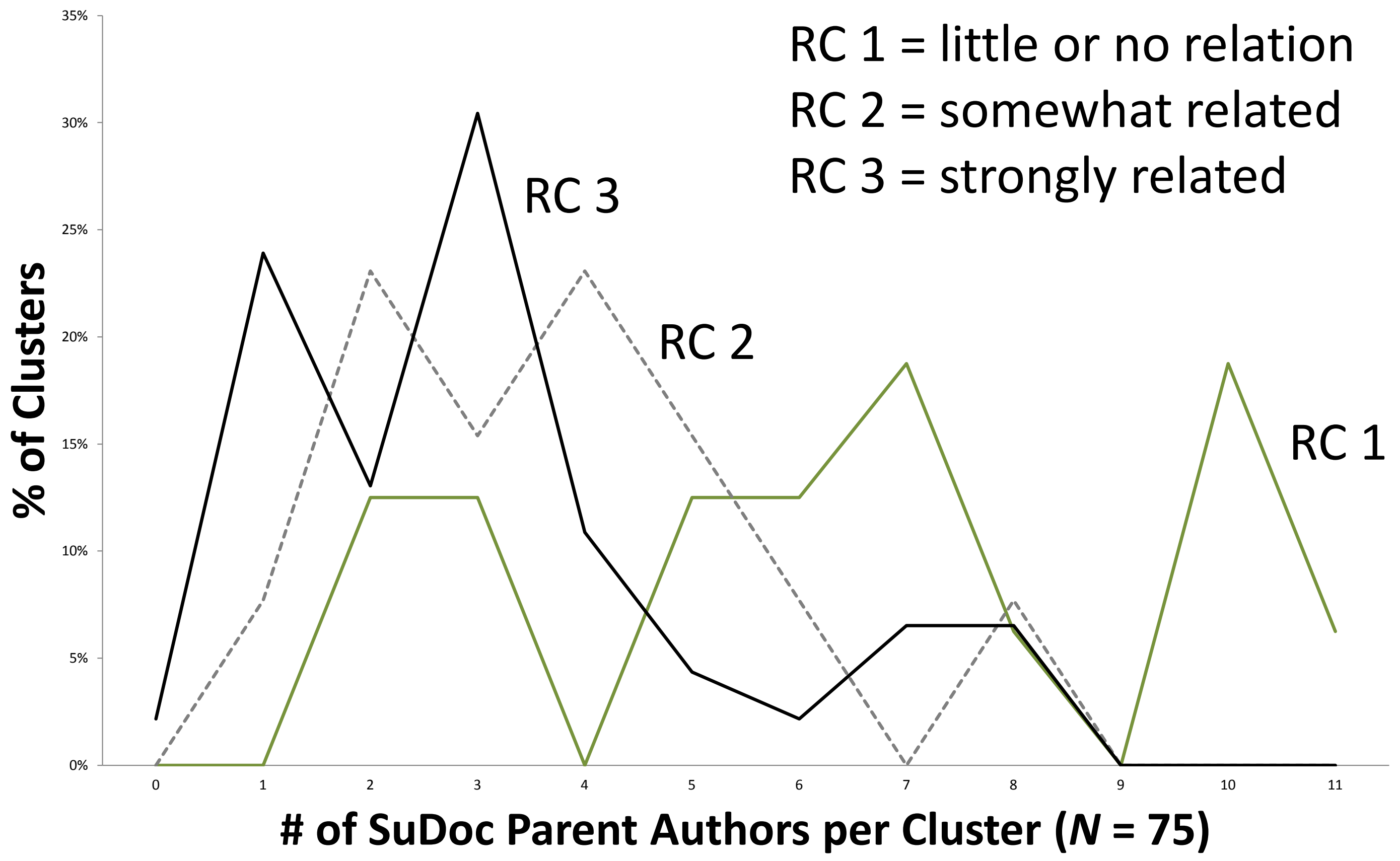
	55-Cluster Set	75-Cluster Set
Identical Clusters	39	39
Unique Clusters	16	36

Website Classification



# Parent Authors	55-Cluster Set	75-Cluster Set
1	16%	16%
≤ 2	27%	32%
≤ 4	60%	67%

Topical Relatedness



Results

- Best clustering result with Linlog Coordinates with Agglomerative Hierarchical Clustering
- Overall: SuDocs Scheme worked well
  - Assigned SuDocs authors to 1,040 subdomains
- Cluster Analysis successfully identified strongly related subject content in the subdomains of 61% of clusters

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

- Identical clusters had the highest percentage of topically related subdomains (72%)
- Unique clusters had a substantially higher percentage of topically related subdomains after subdivision (64% v. 44%)

Conclusions

- Cluster analysis holds promise for organizing Web archives into topically related groupings
- Involving SMEs in limited-scope classification activities may generate meaningful descriptive metadata for resources in focused Web archives
- 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?

Acknowledgements

- EOT Project Partners: Library of Congress, the Internet Archive, the University of North Texas (UNT) Libraries, the California Digital Library, and the US Government Printing Office
- Institute of Museum and Library Services; 2009 -2012 National Leadership Grant

UNT Libraries