Exploratory Analysis of the End of Term Web Archive: Comparing two collections.

Mark Phillips, Dan Chudnov, James Jacobs

June 22, 2016
Web Archiving and Digital Libraries
JCDL 2016

Presidential End of Term 2008 and 2012

EOT2008, EOT2012

EOT2008

Partnership between five institutions to document the transition of the executive branch of government

EOT2012

Provided an event for partners to create a snapshot of the federal domain (.gov & .mil)

EOT2016

Again there will be a transition in the executive branch.

Also four years later, we should do another snapshot.

As we prepare for EOT2016, we've heard a variation of one statement.

What did you all learn in 2008 and 2012 that you will apply to 2016?

The sad truth is that other than organizational optimizations, we don't have much knowledge gained from the two projects.

We've not even compared the two collections at a high level to see where they are similar or different.

This project was a first step in comparing the two EOT archives to better understand them as well as to offer some guidance as we move into the 2016 EOT cycle

Data

CDX files were used for all analysis

UNT Libraries copies of EOT2008 and EOT2012 datasets

EOT2008 – 160,212,141 URIs
EOT2012 – 194,066,940 URIs

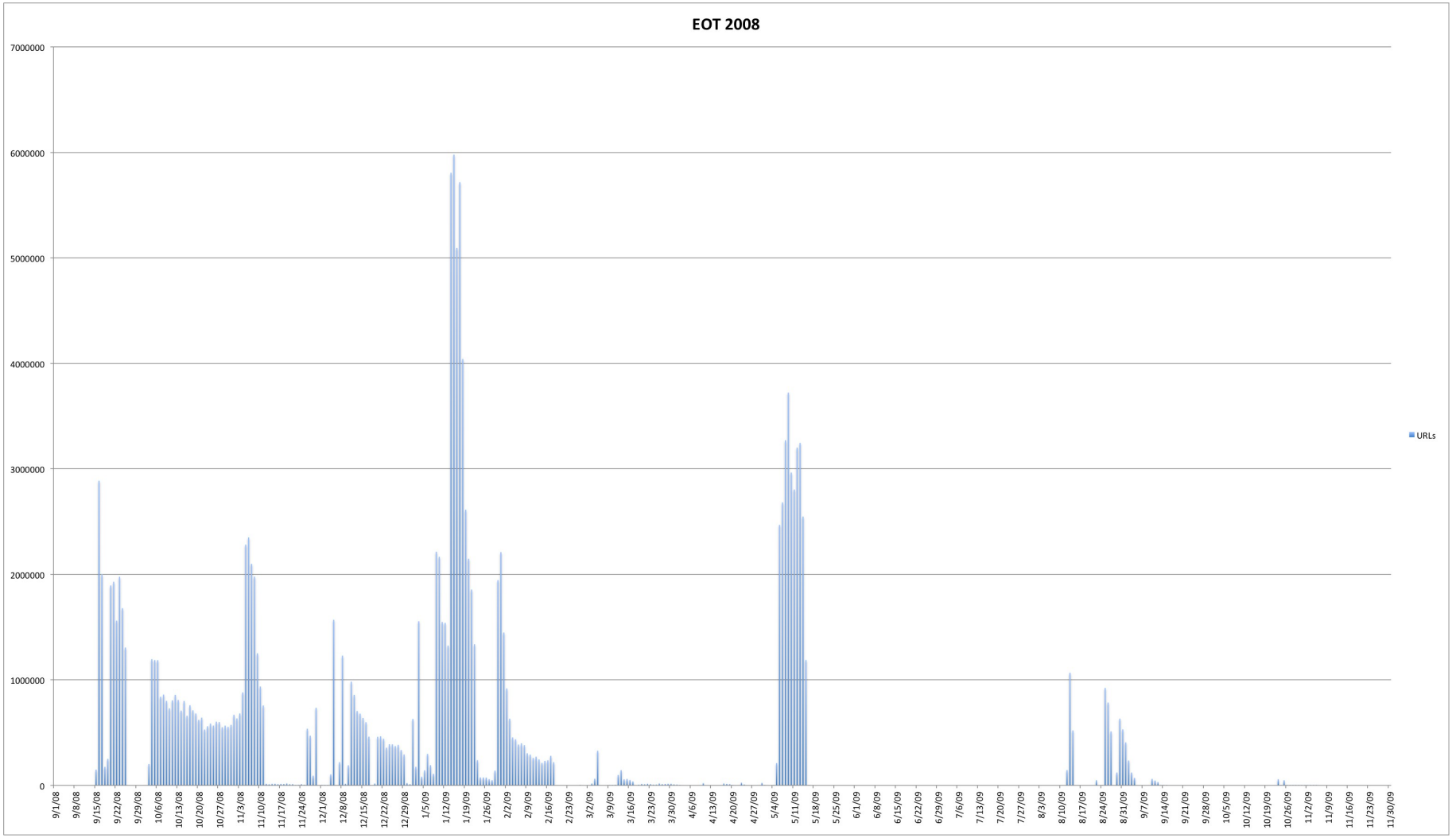
When

The first thing we looked at was when was the content harvested.

In 2008 we had four institutions that harvested.
IA, CDL, LOC and UNT

We had a goal of a snapshot
Before the election
After the election
After the inauguration

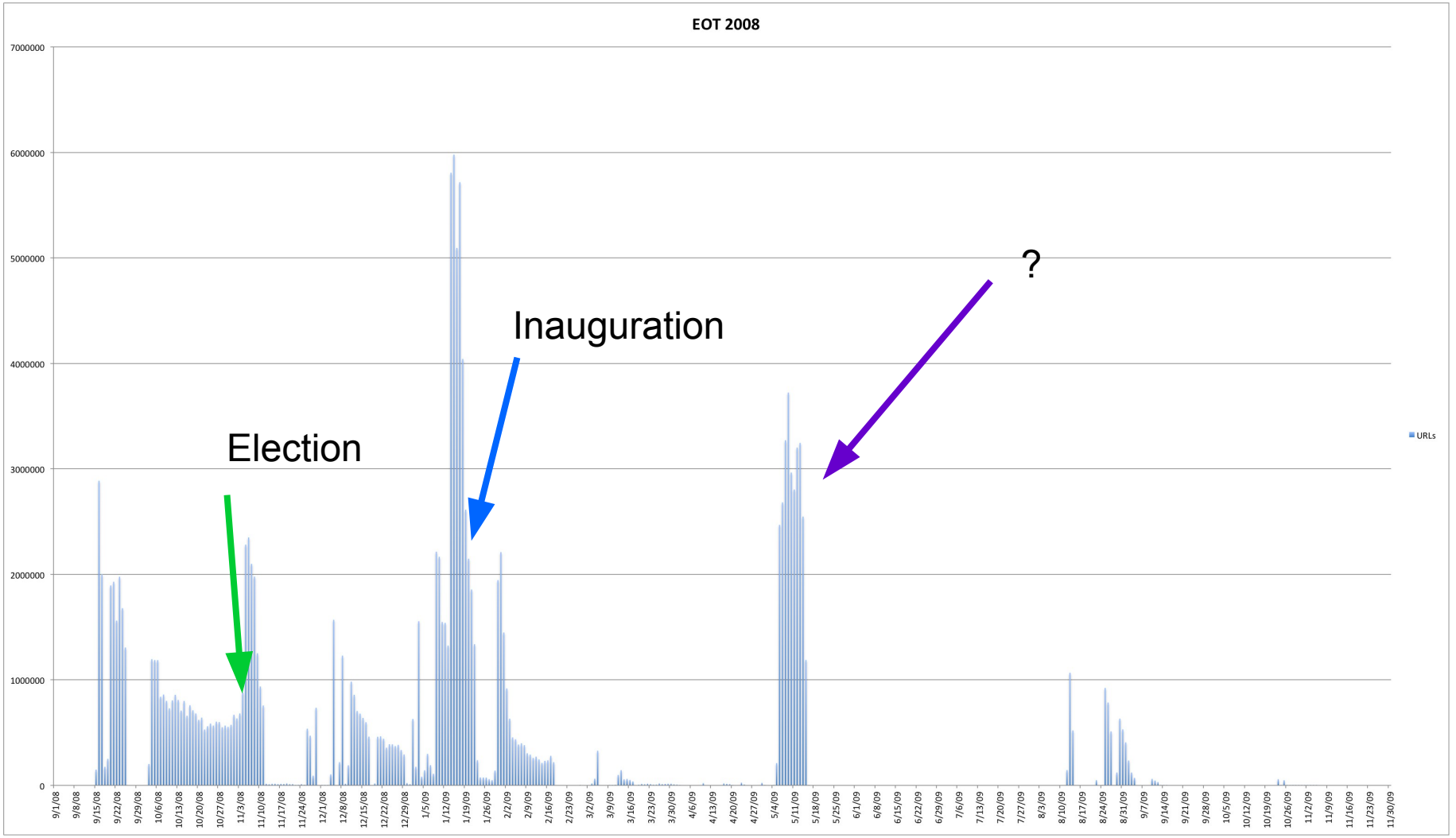
EOT2008 – URLs captured



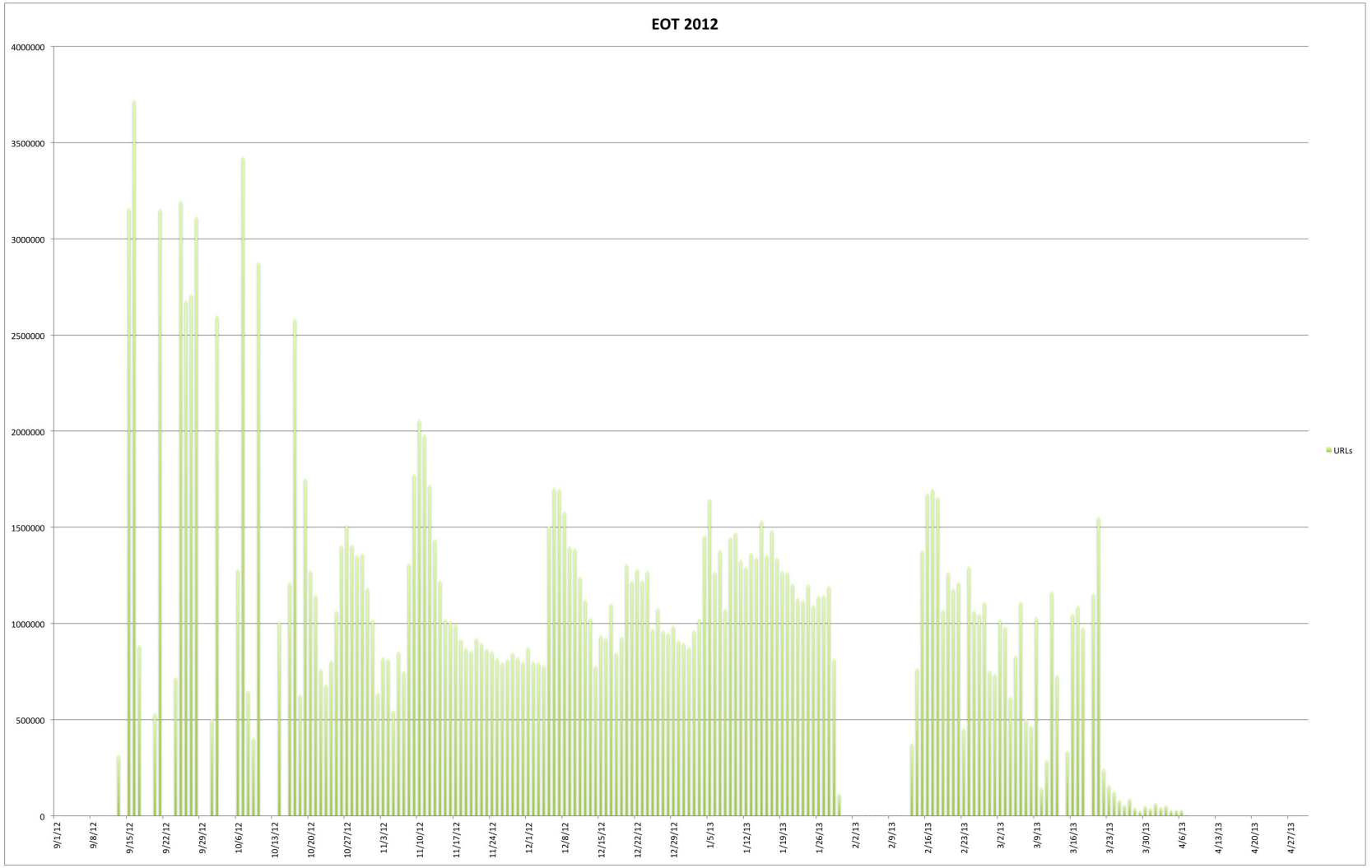
First day of harvesting

September 15, 2008

EOT2008 – URLs captured



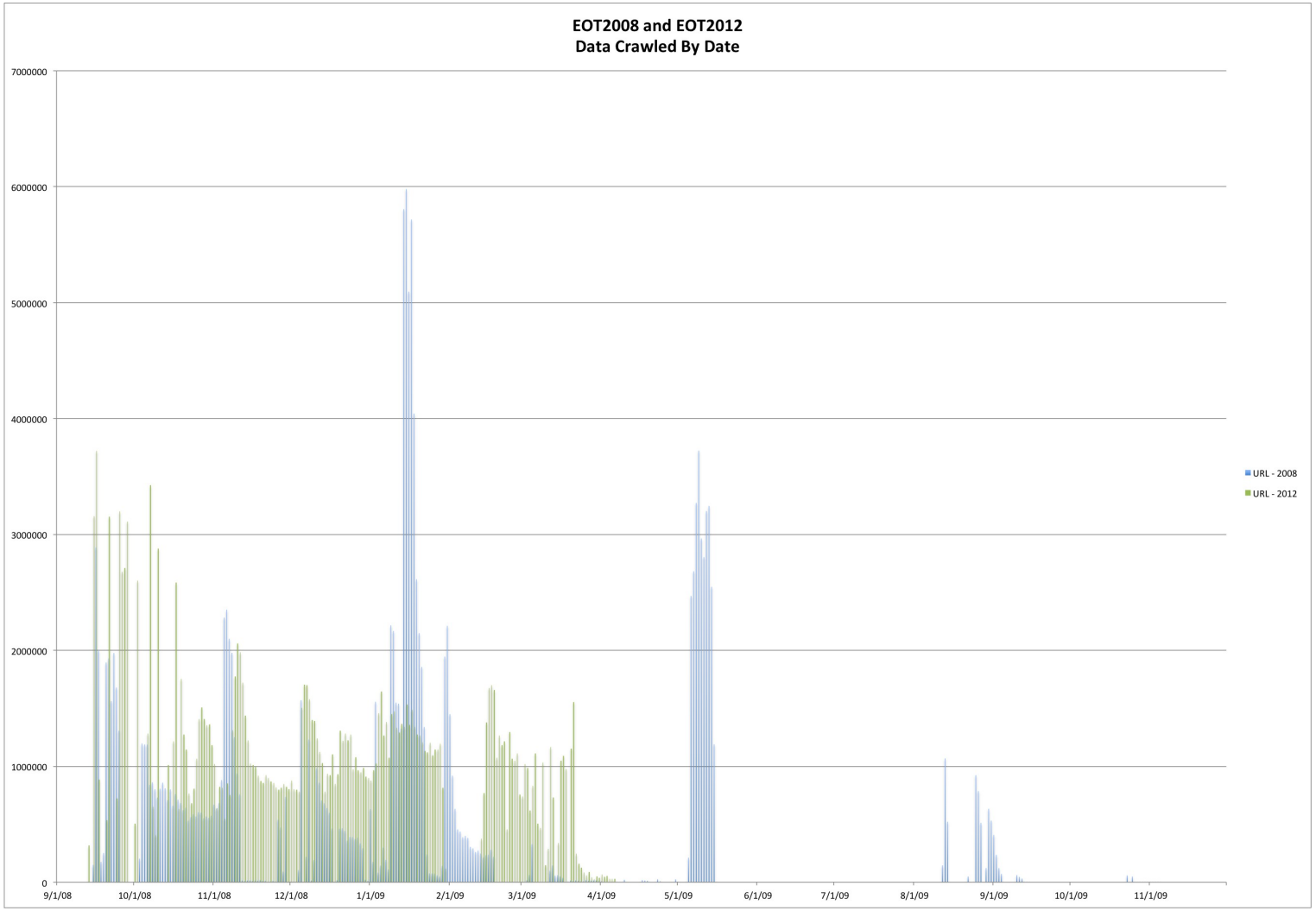
EOT2012 – URLs captured



First day of harvesting

September 13, 2012

EOT2008 and EOT2012

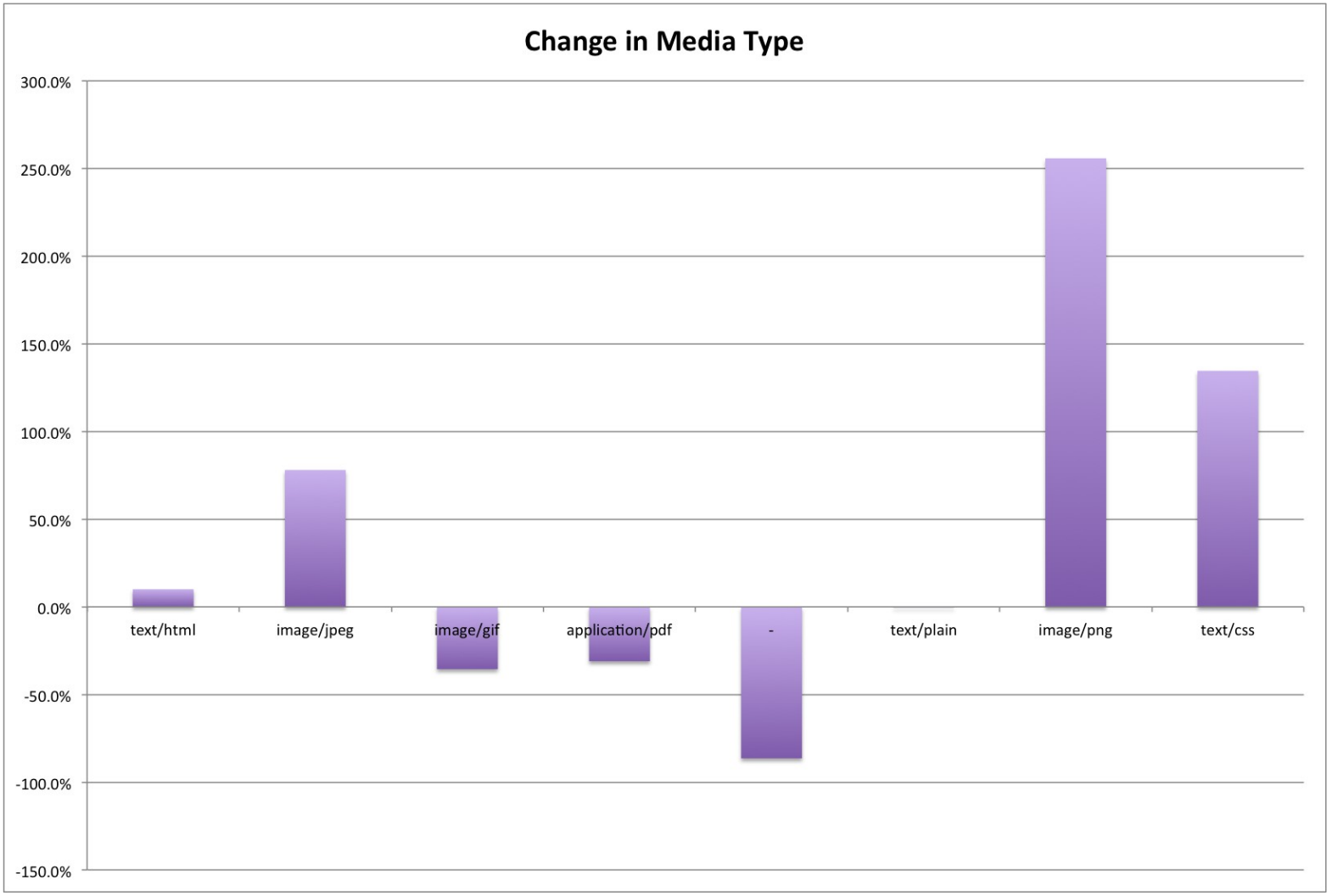


Visually different patterns of crawling.

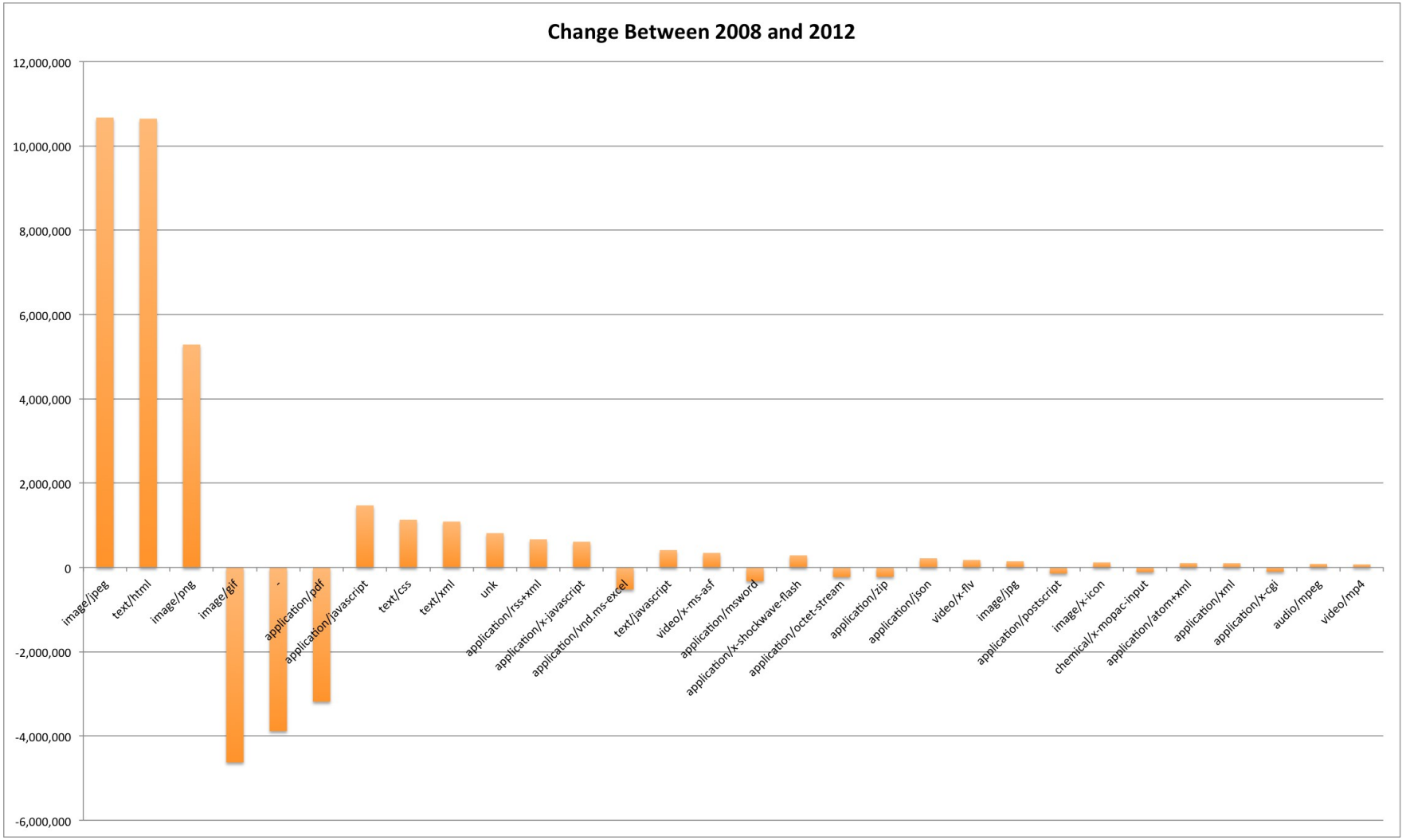
What

Looking at the Media Types between the archives

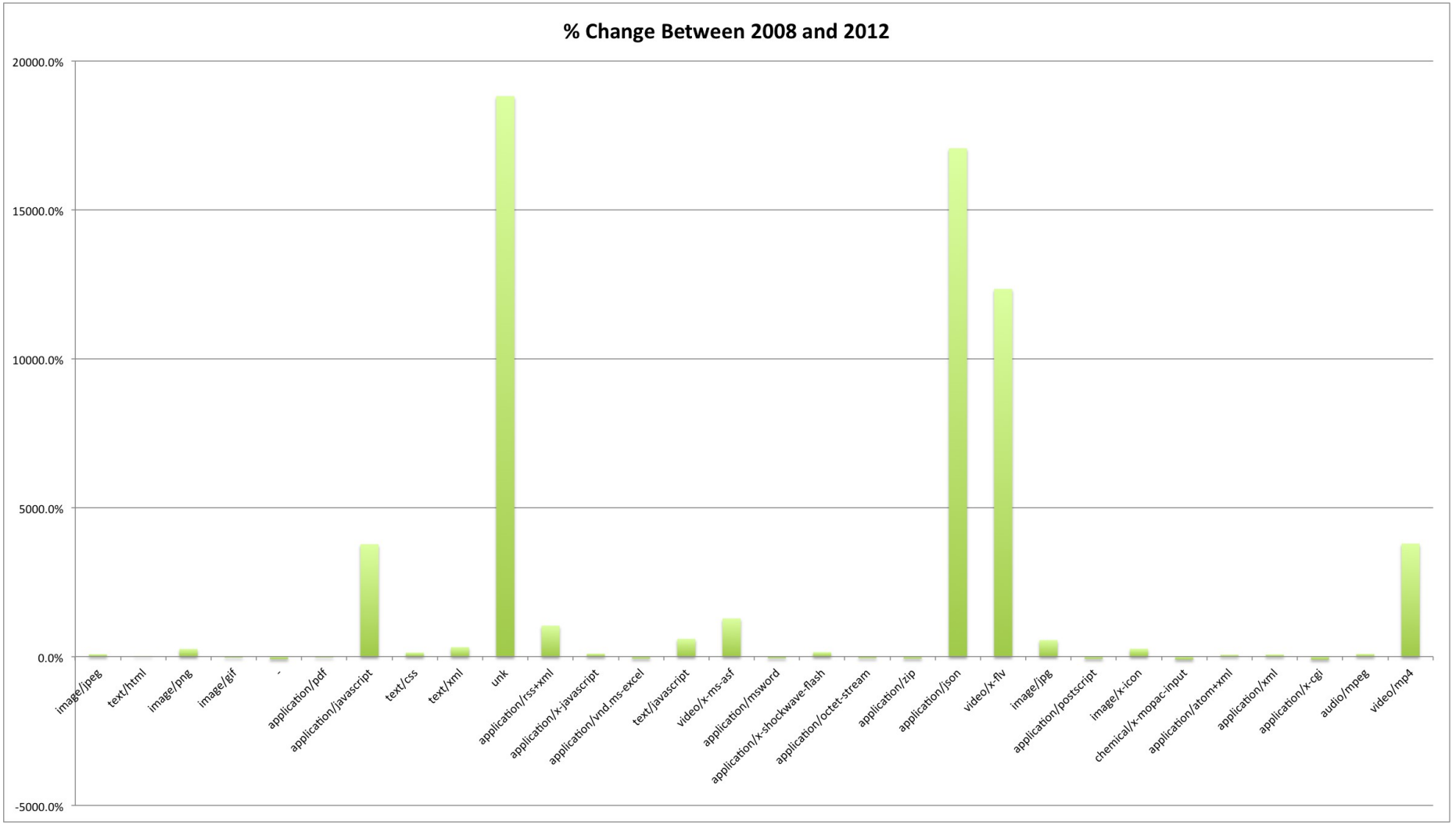
Media Type change from EOT2008 to EOT2012



Top changing Media Types - 2008-2012



Top changing Media Types – 2008-2012 Percentage Change



Where

TLDs, Domains, Subdomains

TLDs

EOT2008 – 241 unique TLDs

EOT2012 – 251 unique TLDs

225 common TLDs

Domains

EOT2008 – 87,889 unique domains

EOT2012 – 186,214 unique domains

30,066 common domains

Subdomains

EOT2008 – 140,939 unique subdomains

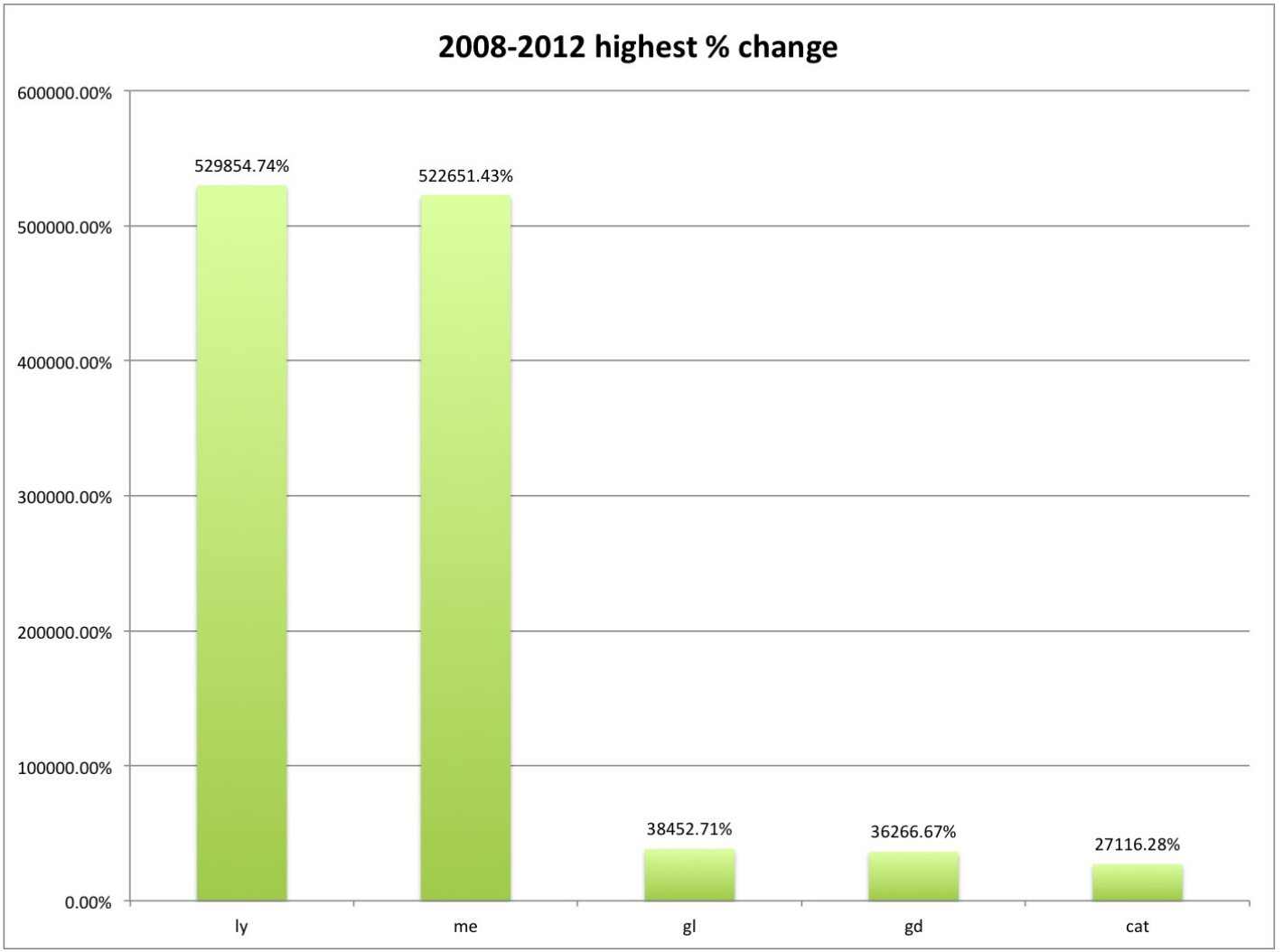
EOT2012 – 352,679 unique subdomains

50,155 common subdomains

EOT2008-EOT2012 – TLD biggest change

TLD	eot2008	eot2012	Change	% change
com	7,809,711	45,594,482	37,784,771	484%
gov	137,829,050	109,141,353	-28,687,697	-21%
mil	3,555,425	16,223,861	12,668,436	356%
net	653,187	9,269,406	8,616,219	1,319%
edu	3,552,509	2,442,626	-1,109,883	-31%
int	135,939	685,168	549,229	404%
uk	70,262	594,020	523,758	745%
ly	95	503,457	503,362	529,855%
org	5,108,645	5,588,750	480,105	9%
us	840,516	474,156	-366,360	-44%

Largest change by percent

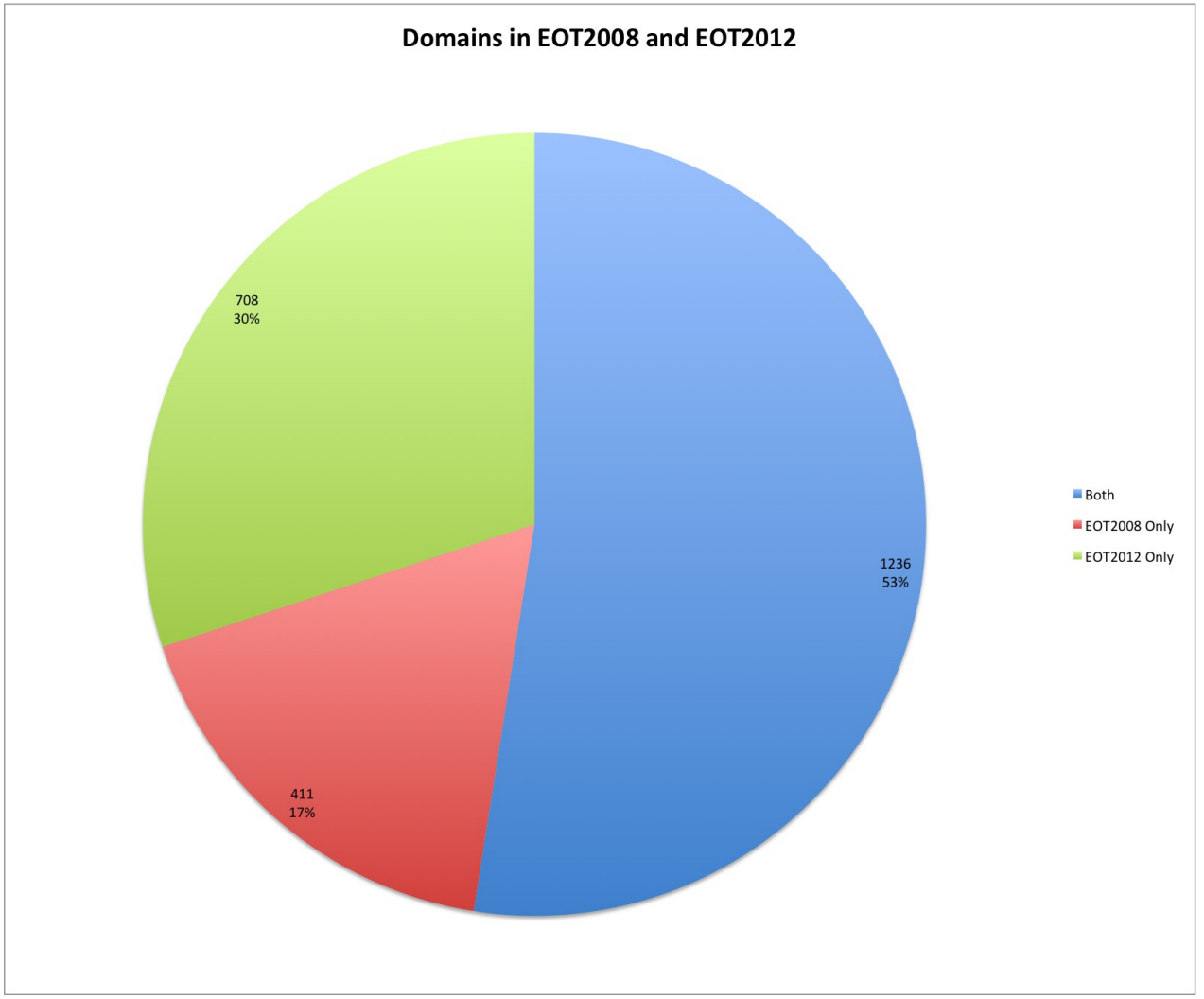


.gov & .mil biggest change - 2008-2012

Domain	EOT2008	EOT2012	Change	% Change
house.gov	13,694,187	35,894,356	22,200,169	162%
senate.gov	5,043,974	9,924,917	4,880,943	97%
gpo.gov	8,705,511	3,888,645	-4,816,866	-55%
nih.gov	5,276,262	1,267,764	-4,008,498	-76%
nasa.gov	6,693,542	3,063,382	-3,630,160	-54%
navy.mil	94,081	3,611,722	3,517,641	3,739%
usgs.gov	4,896,493	1,690,295	-3,206,198	-65%
loc.gov	5,059,848	7,587,179	2,527,331	50%
hhs.gov	2,361,866	366,024	-1,995,842	-85%
osd.mil	180,046	2,111,791	1,931,745	1,073%
af.mil	230,920	2,067,812	1,836,892	795%
ed.gov	2,334,548	510,413	-1,824,135	-78%
lanl.gov	2,081,275	309,007	-1,772,268	-85%
usda.gov	2,892,923	1,324,049	-1,568,874	-54%
congress.gov	1,554,199	40,338	-1,513,861	-97%

What wasn't there?

.gov and .mil domains in EOT2008 and EOT2012



Top 30 .gov & .mil domains present in 2008 missing in 2012

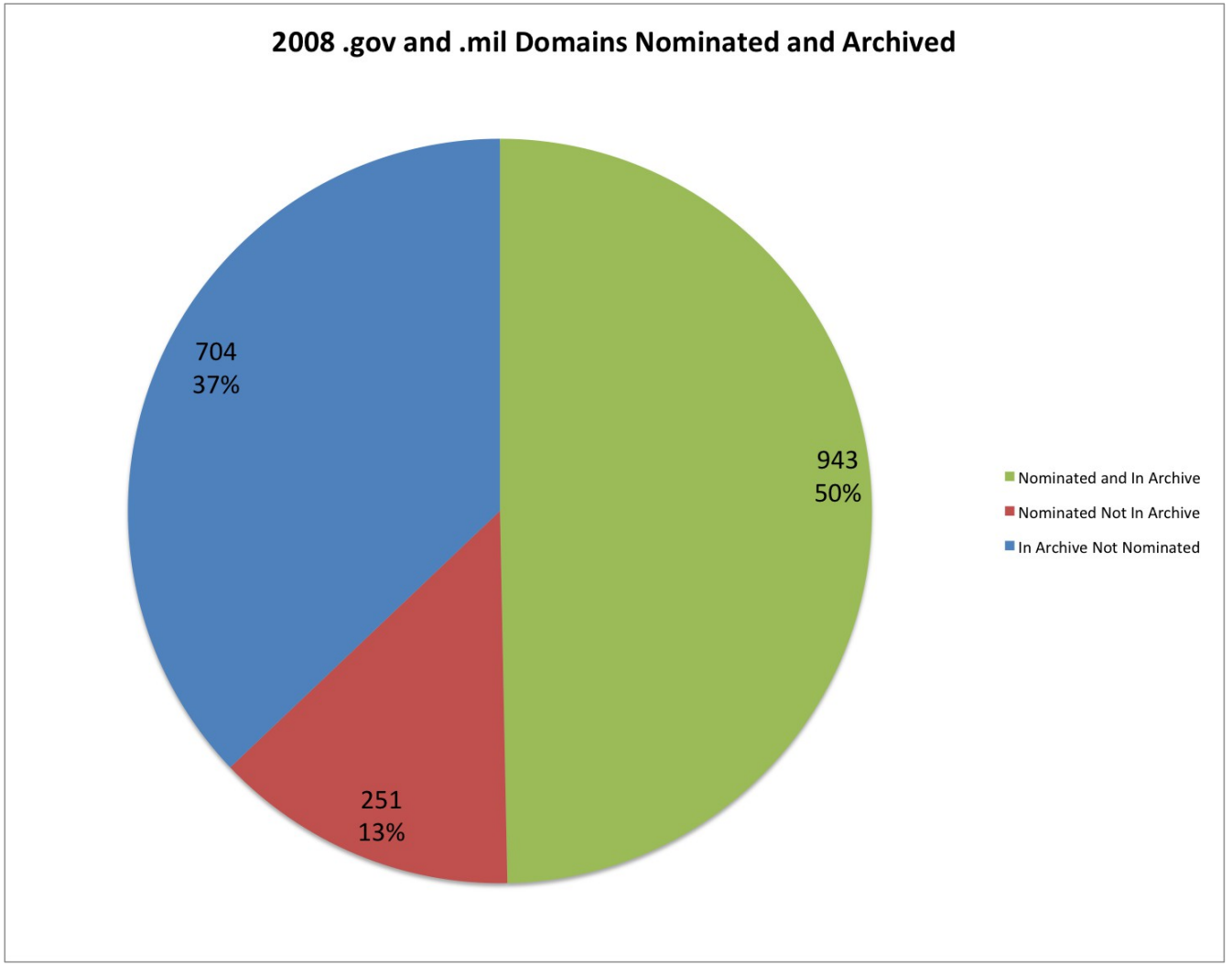
Domain	URLS	Domain	URLs
geodata.gov	812,524	egrpra.gov	54,775
nifl.gov	504,910	4women.gov	45,684
stat-usa.gov	398,961	4woman.gov	42,192
tradestatsexpress.gov	243,729	nypa.gov	36,099
arnet.gov	174,057	nhmfl.gov	27,569
acqnet.gov	171,493	darpa.gov	21,454
dccourts.gov	161,289	usafreedomcorps.gov	18,001
web-services.gov	137,202	peacecore.gov	17,744
metrokc.gov	132,210	californiadesert.gov	15,172
sdi.gov	91,887	arpa.gov	15,093
davie-fl.gov	88,123	okgeosurvey1.gov	14,595
belmont.gov	87,332	omhrc.gov	14,594
aftac.gov	84,507	usafreedomcorp.gov	14,298
careervoyages.gov	57,192	uscva.gov	13,627
women-21.gov	56,255	odci.gov	12,920

Top 30 .gov & .mil domains new in 2012

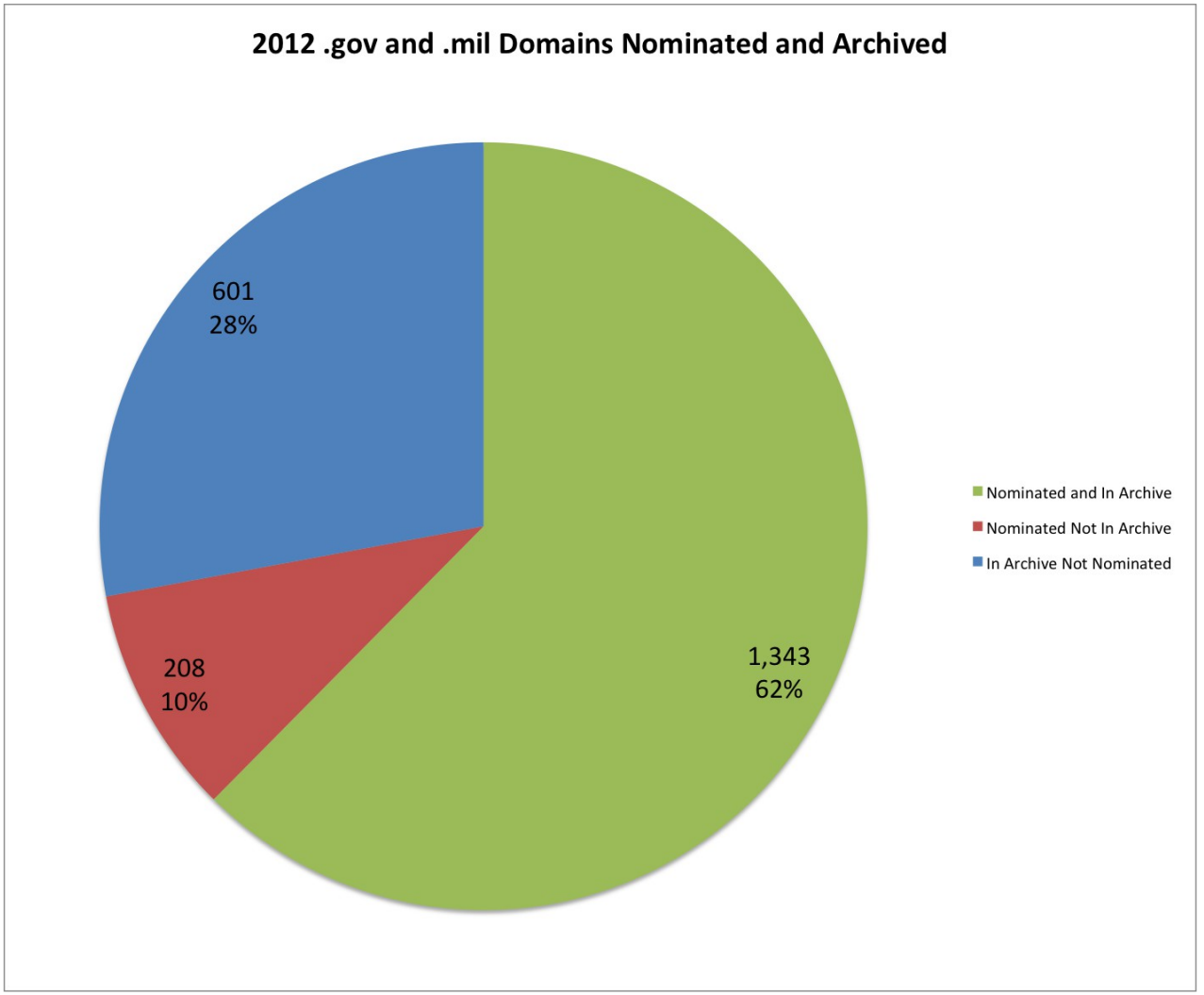
Domain	URLS	Domain	URLs
militaryonesource.mil	859,843	broadbandmap.gov	72,889
consumerfinance.gov	237,361	saferproducts.gov	65,387
nrd.gov	194,215	challenge.gov	63,808
wh.gov	179,233	healthdata.gov	63,105
pnnl.gov	132,994	marinecadastre.gov	62,882
eia.gov	112,034	fatherhood.gov	62,132
transparency.gov	109,039	edpubs.gov	58,356
nationalguard.mil	108,854	transportationresearch.gov	58,235
acus.gov	93,810	cbca.gov	56,043
404.gov	82,409	usbonds.gov	55,102
savingsbondwizard.gov	76,867	usbond.gov	54,847
treasuryhunt.gov	76,394	phe.gov	53,626
fedshirevets.gov	75,529	ussavingsbond.gov	53,563
onrr.gov	75,484	scienceeducation.gov	53,468
veterans.gov	75,350	mda.gov	53,010

Curator Intent?

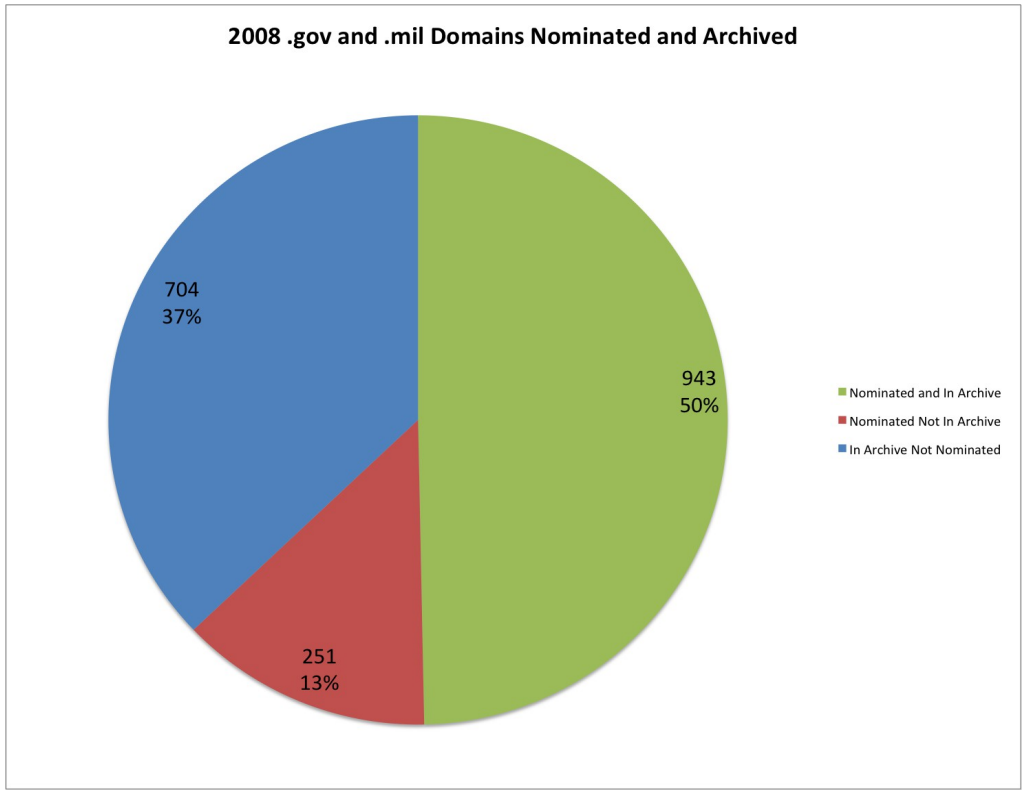
2008 .gov and .mil Domains Nominated and Archived

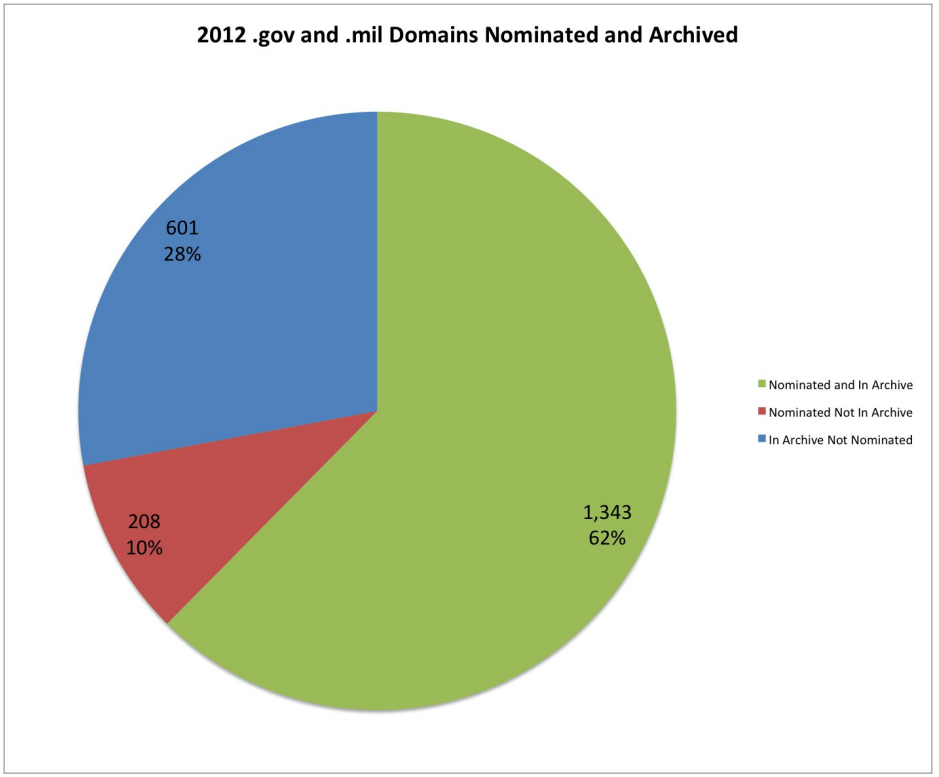


2012 .gov and .mil Domains Nominated and Archived



2008 and 2012 .gov and .mil domains nominated vs archived





As I mentioned at the beginning, this is a first attempt at comparing these two collections

There are a number of other questions that we want to answer

Some can be answered with just CDX files

Others will require the arc/warc data and content analysis

All data and code is available here -

https://github.com/vphill/eot-cdx-analysis

Questions?

mark.phillips@unt.edu