Content of Municipal Government Websites

Markus Neumann Bruce Desmarais Hanna Wallach

March 1, 2017

Abstract

We study the content of municipal government websites....

1 Introduction

2 Data

The General Services Administration (GSA) maintains all .gov addresses, and provides a complete[1] list of all such domains to the public through GitHub[2]. This list is updated once per month - we rely on the version released on January 16, 2017. The data from the GSA contains the following variables: One, domain name, specifically, the all-uppercase version of domain and top-level domain (for example, 'ABERDEENMD.GOV'). Two, the type of government entity to which the domain is registered, such as city, county, federal agency, etc. Three, for federal agencies, the name is specfied. Finally, the city in which the domain is registered, is noted.

Here, we focus only on cities. As a first step, we use a webdriver-controlled browser (Firefox/Selenium/Geckodriver) to test whether all of the city websites actually work. Of the 2425 domains listed by the GSA as cities, 292 are not accessible. Furthermore, the .gov domain, as registered at the GSA, is frequently not the website a city actually uses. In many cases, these sites redirect to another address, sometimes not a .gov domain (in this case, we simply use this domain). We record these URLs, as they are required to retrieve the images websites stored in the Wayback Machine (WbM).

In order to provide an overview of our coverage (as not all cities, towns and villages use .gov addresses), we merge this list with U.S. Census data[3]. Here, several limitations in the GSA data need to be accounted for: One, even though the GSA nominally separates websites of cities and counties, some of the domains categorized as cities actually belong to counties. The same is true for townships and boroughs. Ergo, we eliminate all websites belonging to these three types of entities by hand. Furthermore, the city name, as given by the GSA, refers to the city in which the domain is registered, which is not necessarily equivalent to the city the website serves. In many cases, a website of a larger city may be registered to one of its subdivisions (for example, the website of New York is registered to Brooklyn), or vice versa (for example, the website of Homecroftin, a small town within Indianapolis, is registered to the city as a whole). Consequently

[1] Domains used for testing and internal programs are excluded.

[2] https://github.com/GSA/data/tree/gh-pages/dotgov-domains

[3] http://www2.census.gov/programs-surveys/popest/datasets/2010-2015/cities/totals/sub-est2015_all.csv

we fix mismatches between websites and cities manually. Finally, a number of cities are simply misspelled, which we also correct by hand.

After the counties, townships and cities that cannot be matched to the Census data[4] and duplicate websites (some cities have more than one website) are removed, 1813 domains/cities remain.

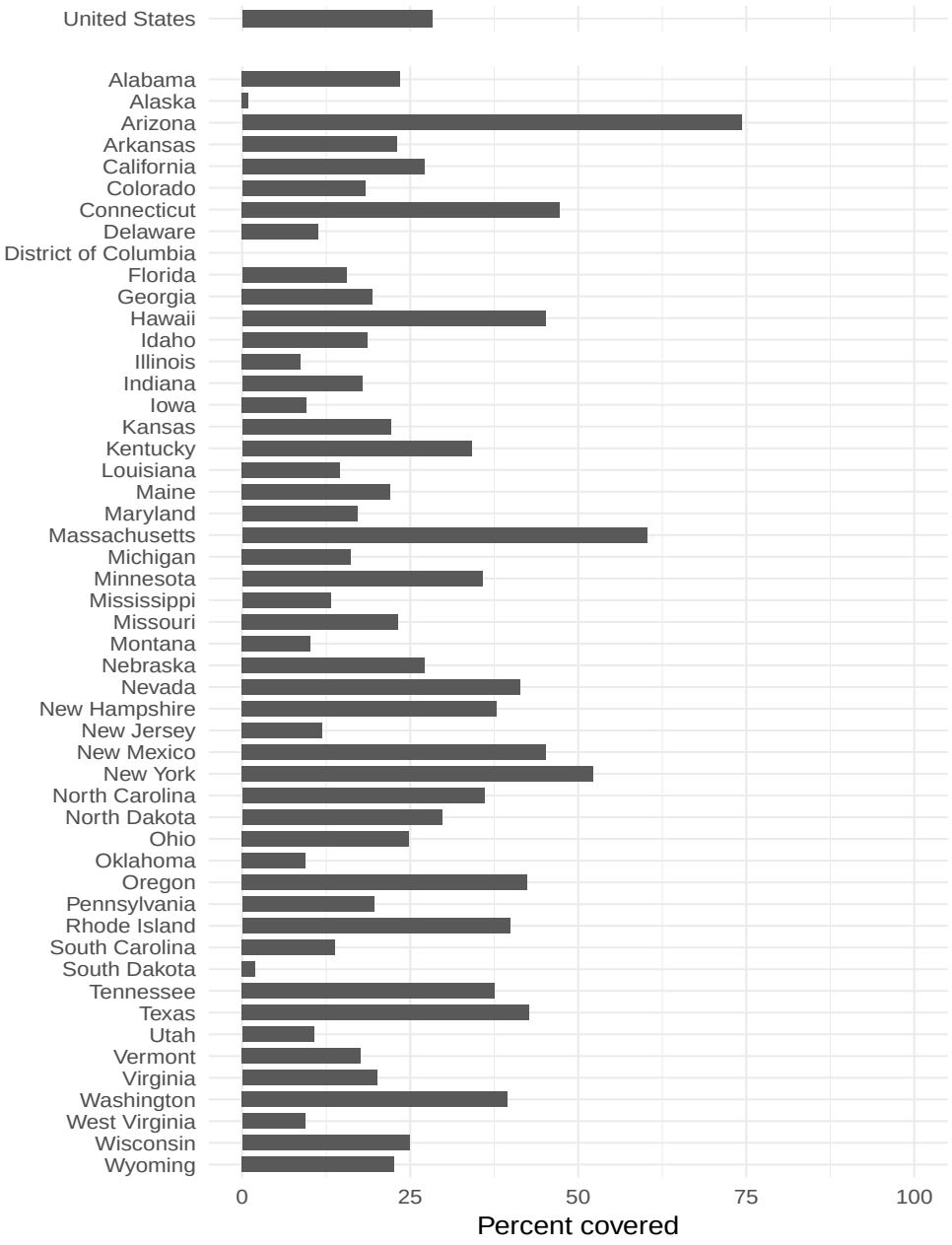
These cities contain 90,616,865 people, and thus about 28% of the U.S. population (see figure 1).

We use the resulting list of websites to acccess their copies stored in the Internet Archive's Wayback Machine. To this end, we rely on the Ruby Gem 'Wayback Machine Downloader'[5] (WbMD). We supply the URL that each .gov website redirects to to the WbMD, which then downloads every file present in the WbM from a snapshot in October 2016, or, if not available, as soon as possible after this point.

<Note: We have not actually done this last step for all websites (however, the R script which runs the Ruby package is already set up to do so once we need to). Instead 10 websites were randomly sampled from an older version of the GSA list, which still contained counties and townships, which is why one of the 10 websites is from Dutchess County, NY.>

[4]There are five cities that are not contained in the Census data

[5]https://github.com/hartator/wayback-machine-downloader


United States
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
0
25
50
75
100
Percent covered

File type	Occurrences
.pdf	1371
.html	819
.png	210
.jpg	131
.gif	99
.js	51
.PDF	50
.aspx	43
.doc	35
.css	32
.JPG	26
.Net	12
.xlsx	6
.docx	5
.ttf	3
.xml	3
.htm	2
.woff	2
.xls	2
.asp	1
.eot	1
.GIF	1
.ico	1
.PNG	1
.ppt	1
.swf	1
.txt	1

Table 1: File types in scraped websites

Website	Files	Size (MB)
brownsvilletn.gov	188	14328
www.centralpointoregon.gov	150	137440
www.dedham-ma.gov	603	212572
www.duncanok.gov	84	47064
www.dutchessny.gov	110	291376
www.ennistx.gov	200	26244
www.greenvillenc.gov	333	25732
www.romi.gov	491	112584
www.trumbull-ct.gov	787	191540
www.westonct.gov	861	213140

Table 2: Test websites

	tf
tax	176324
date	98949
due	97192
amt	96382
town	86726
value	81119
total	70825
parcel	63589
county	56201
market	51758
east	51357
full	51199
nrth	50566
book	50306
deed	50231
bill	49719
acres	48935
acct	44871
csd	43792
owners	43510
res	41803
family	36883
fire	35156
school	33685
name	30382
red	30362
taxable	30248
hook	29902
homestead	29287
outside	28593

Table 3: Top term frequencies for 10 test websites