Government Websites As Data: Understanding how Mayoral Partisanship Shapes Municipal Website Content

January 17, 2020

Abstract

A local government's website is an important source of information about policies and procedures for residents and community stakeholders . Existing research in political science and related fields has relied on manual methods of website content collection and processing, limiting the scale and scope of website content analysis. In this research note, we propose that the automated collection of website content from large samples of government websites can can compliment more targeted manual methods, and offer contributions through comparative analyses. Our approach, which represents a considerable improvement in scalability, involves downloading the entire contents of a website, extracting the text and discarding redundant information. In our application, we collect and analyze a new and innovative dataset—the websites of over two hundred municipal governments in the United States. We build upon recent research that analyzes how variation in the partisan control of government relates to content made available on the government's website. Using topic modeling methods, we find that cities with Democratic mayors provide more information on policy deliberation and crime control, whereas Republicans prioritize basic utilities and services such as water, electricity, and fire safety. As an additional contribution, we provide an R package that can be used to process government website files into raw text.

Government Websites As Data: Understanding how Mayoral Partisanship Shapes Municipal Website Content

1 Introduction

Government websites convey voluminous information about all aspects of government policy priorities, policy implementation, and public deliberation. The vital role of official websites in connecting the government and the governed has motivated a wave of research on the contents of government websites, focusing in particular on textual contents (e.g., Grimmelikhuijsen 2010; Wang et al. 2005; Osman et al. 2014). The conventional approach to data collection in projects focused on government websites involves manual content extraction from each website in the dataset. Though accurate, the manual approach to data collection is costly for large-scale analysis. We propose the use of automatic web-scraping to gather and process government websites in order to build datasets that can be used for text analysis, as well as the solutions we adopt. We apply this approach to build a novel dataset of U.S. municipal government website contents, and analyze how the textual contents of city government websites in six American states correlate with the partisanship of the city mayor.

Our contributions are two-fold. First, we propose, and provide software to apply, the large-scale automated collection of textual data from government websites–covering entire contents including the plain HTML files, and linked files in various formats (e.g., DOC, PDF, and TXT) into plain text for analysis. Second, we gather and analyze a dataset that covers the textual contents of websites from over two hundred municipal governments in the United States. By studying the covariation of topical contents on these websites with the partisanship of the city mayors, we both illustrate the utility of automated web content collection from government websites, and present new findings on the relationship between government website concepts and executive partisanship.

2 Politics and Government Website Content

Though government websites serve largely instrumental service-delivery purposes, they also offer officials a prime venue via which to communicate policy goals and accomplishments, which inevitably reflect officials' politics. In the current paper, we focus on the running example of the reflection of mayoral partisanship on municipal government websites. A substantial body of research has found that the partisanship of the mayor affects city governance along multiple dimensions of spending and policy attention (e.g., Gerber and Hopkins 2011; de Benedictis-Kessner and Warshaw 2016; Einstein and Glick 2016). Official city websites allow mayors to present their views and policy priorities to the public. In local politics, where campaign funds are low, this lends incumbents a crucial advantage in becoming more well-known among their constituencies (Stanyer 2008). Local government websites are frequently visited by the public (Thomas and Streib 2003). City websites can be used to communicate the stance of a mayor on social or economic programs. Consider the screenshot of the Gary, Indiana homepage in Figure 1. This provides a clear example of the utility of a city website for communicating the mayor's policy priorities and accomplishments.

[Figure 1 about here.]

The existing research that uses scraped websites provides an indication of the theoretical value of empirical analysis of web contents. Research on 'e-governance' evaluates government websites in terms of accessibility, ease-of-use, and function (e.g., McNeal et al. 2003; Tolbert et al. 2008; McNutt 2010). As an example, Grimmelikhuijsen and Welch (2012) study local government websites of Dutch municipalities to measure government transparency regarding air quality in the municipalities. The websites of politicians and their parties have also been the object of research (Druckman et al. 2009, 2010; Esterling et al. 2011; Esterling and Neblo 2011; Norris 2003). For example, Druckman et al. (2010) analyze the issues engaged on websites for candidates in U.S. Congressional elections, and find that candidates strategically engage just a few issues based on the priorities in their districts and the characteristics of their opponents.

3 Data: US Municipal Government Website Text

For data availability reasons when it comes to mayoral partisanship, we focus our analysis of municipal websites on six states—Indiana, Louisiana, New York, Washington, California, and Texas. The websites were scraped in March 2018. The selection of states and cities is largely dictated by the presence of partisan mayors. Municipal elections in Indiana and Louisiana are partisan across the board, so our sample is primarily focused on these two states. For Indiana and Louisiana, all cities with a website are included, resulting in a considerably larger sample than for the other four states. New York and Washington do not have nominally partisan elections, but for a subset of cities, partisanship can be determined through contribution data (see supplementary material for more detail). California and Texas contain a number of large cities whose mayors are sufficiently well-known for their partisanship to be available. Our sample is well-balanced on a number of theoretically important dimensions. One, each of the four Census regions are represented with at least one state. Two, we have a fairly well-balanced sample with respect to the urban/rural cleavage. Furthermore, the sample is politically balanced—we have three blue states, and three red states. The partisan breakdown of city websites by state is provided in the supplementary material. This dataset of city website contents represents a contribution in the growing area of cross-municipality datasets covering local governements (e.g., Marschall and Shah 2013; Sumner, Farris, and Holman Sumner et al.). Details on the sources and methods of raw data collection can be found in the supplementary material.

4 Partisan Language on Municipal Websites

City mayors use government websites to present their policy priorities to the public. Consider an example; (Formicola et al. 2003, p.55) document a significant website content change during a transition in mayoral administrations in the city of Indianapolis. Under the Republican mayor Stephen Goldsmith, voluminous content was added to the city website in connection with the Front Porch Alliance—a faith-based initiative to create partnerships with religious organizations for the use and administration of city resources. Faith-based initiatives represent a type of public-private partnership that is popular with Republicans (Saperstein 2003). When Democrat Bart Peterson took office in 2000, the material related to the Front Porch Alliance was removed from the website. We consider whether the partisan manipulation of city website contents documented in this example holds in a large-scale and more recent sample of city websites. We illustrate the analysis of municipal website content by studying how differences in website content correlate with the partisanship of the city's mayor. As we reviewed above, the partisanship of the mayor has been found in past research to affect several features of city governance. However, Gerber and Hopkins (2011) note that, due to the constraints of state and national policies, municipalities lack discretion in many domains of governance. These constraints do not apply to website contents. City governments have great discretion in composing their websites, modifying website content is low cost relative to other policy changes, and city websites provide an effective and often-used means of communication with city residents.

We use well established web-scraping and text-processing tools to go from a list of municipal website URLs to a corpus of clean text to analyze. The first step is to automatically download the entire file contents of municipal websites, which we do using the Unix tool `wget` (Glez-Peña et al. 2013). Step two is to convert each file (including, e.g., HTML, DOC, PDF) into a plain text file, which we do using the `readtext` R package (Benoit and Obeng 2019). Step three is to discard website boilerplate text (e.g., navigation menus), which we do using the "boilerpipe" algorithm (Kohlschütter et al. 2010), which is implemented in the R package `boilerpipeR`. The fourth step is a final processing wave applied to the resulting text corpus, which includes, e.g., removing strings that are not English words. This web-scraping and processing pipeline is described in greater detail in the supplementary material, and we have implemented all but the first step in a new R package (`gov2test`).

To study content differences between government websites based on mayoral partisanship, we draw upon a recently-developed model for text, the structural topic model (STM), developed by Roberts et al. (2014). Building on the conception of "topics" in conventional topic models (Valdez et al. 2018), in the STM a topic is a multinomial distribution defined on the word types in the corpus dictionary. The log-odds of the topic probabilities in each document-specific multinomial distribution over topics are drawn from a multivariate normal distribution in which the topic-specific means are determined by a linear regression function that associates document-attributed covariates with topics. For example, in the context of municipal website content, the structural topic model can be used to estimate a regression coefficient that defines the linear relationship between the log-odds of the municipality's population and the log-odds of each topic. For our primary empirical investigation, the STM provides a tool to estimate the relationship between the party of the city's mayor and the prevalence of each topic. We also include the municipality's population and median income as covariates. Further details on and results from our STM specification can be found in the supplementary material.

4.0.1 Structural topic model results

The results are shown in Table 1. First, it is notable that the 95% credible interval includes zero (indicated by rows with white backgrounds) in only seven of the sixty topics. This suggests that the topics discussed on city websites varies systematically with the partisanship of the mayor. Many of the topics associated with Democrats fit with what we understand to be national party priorities. Topic **52**, on affordable housing, clearly resonates with the Democratic party's appeal to low-income voters. Topic **6** ('race', 'islander', 'census, 'female') covers racial and gender identity issues. Similarly, employee rights and benefits are represented in topics **10** and **29**. Democrats also exhibit a strong preference for words related to public finances, such as Topic **58** ('budget', 'revenue', 'expenditure'), Topic **45** ('asset', 'actuarial', 'liability', 'financial'), Topic **35** ('bond', 'obligation', 'proceeds') as well as Topic **55** ('taxable', 'deed', 'value'). We suspect that the association of Democratic mayors with finance-related terms is indicative of a greater willingness to emphasize the city's efforts to raise and spend money, and take credit for those efforts (e.g., the Gary, IN example in Figure 1). This finding is consistent with Einstein and Kogan (2015), who show that Democratic mayors tend to favor greater spending. A second, consistent Democratic focus appears to be law enforcement: The most Democratic topic, **59** ('burglary', 'robbery', 'theft', 'homicide') is clearly focused on crime. On the one hand, Democratic partisans have a more negative perception of the police, rating it considerably more negatively on the appropriate use of force and the equal treatment of minorities (Brown 2017). On the other hand, the literature has also shown that cities with a higher Democratic vote share spend more on law enforcement, even after controlling for crime (Einstein and Kogan 2015).

City websites with Republican mayors, meanwhile, exhibit a pronounced focus on the essential functions of government. Basic utilities such as energy (Topic **20**), fire protection (Topic **51**), vaccination (Topic **2**), and sanitation (Topic **47**), are prevalent among cities with Republican mayors. These basic service issues cannot be found among topics prevalent in cities with Democratic mayors. Similarly, zoning issues figure prominently in the set of republican topic (Topic **19**), which fits with the findings of Sorens (2018) that Republicans are more supportive of restrictive residential zoning rules. The Democratic topics also include one that is somewhat focused on zoning, Topic **39** ('downtown', 'mixed', 'density'), but emphasizes mixed-use zoning—a loosening of conventional single-use zoning rules.

[Table 1 about here.]

5 Conclusion

We have proposed the automatic collection and processing of government websites for comparative content analysis. We have produced an R package `gov2text`, in which we have implemented and wrapped the core components of the methods we use. This methodology holds the potential to vastly scale up the data collection efforts underpinning the growing body of research that is focused on government website analysis. Through an application to the analysis of municipal websites in six different states, we show how our pipeline is capable of gathering corpora that shed light on the forms and functions of local government. We find that government website contents are associated with the partisanship of the mayor in ways that would be expected based on the parties' national priorities and past research on the effects of mayoral partisanship on city governments.

References

Benoit, K. and A. Obeng (2019). *readtext: Import and Handling for Plain and Formatted Text Files*. R package version 0.74.

Brown, A. (2017). Republicans more likely than Democrats to have confidence in police.

de Benedictis-Kessner, J. and C. Warshaw (2016). Mayoral partisanship and municipal fiscal policy. *The Journal of Politics 78*(4), 1124–1138.

Druckman, J. N., C. L. Hennessy, M. J. Kifer, and M. Parkin (2010). Issue Engagement on Congressional Candidate Web Sites, 2002—2006. *Social Science Computer Review 28*(1), 3–23.

Druckman, J. N., M. Kifer, and M. Parkin (2009). Campaign Communications in U.S. Congressional Elections. *American Political Science Review 103*(03), 343–366.

Einstein, K. L. and D. M. Glick (2016). Mayors, partisanship, and redistribution: Evidence directly from us mayors. *Urban Affairs Review*, 1078087416674829.

Einstein, K. L. and V. Kogan (2015). Pushing the City Limits: Policy Responsiveness in Municipal Government. *Urban Affairs Review*, 1–30.

Esterling, K. M., D. M. Lazer, and M. A. Neblo (2011). Representative communication: Web site interactivity and distributional path dependence in the us congress. *Political Communication 28*(4), 409–439.

Esterling, K. M. and M. A. Neblo (2011). Explaining the Diffusion of Representation Practices among Congressional Websites. *Working Paper*, 1–42.

Formicola, J., M. Segers, and P. Weber (2003). *Faith-based Initiatives and the Bush Administration: The Good, the Bad, and the Ugly*. Rowman & Littlefield.

Gerber, E. R. and D. J. Hopkins (2011). When mayors matter: estimating the impact of mayoral partisanship on city policy. *American Journal of Political Science 55*(2), 326–339.

Glez-Peña, D., A. Lourenço, H. López-Fernández, M. Reboiro-Jato, and F. Fdez-Riverola (2013). Web scraping technologies in an api world. *Briefings in bioinformatics 15*(5), 788–797.

Grimmelikhuijsen, S. G. (2010). Transparency of public decision-making: Towards trust in local government? *Policy & Internet 2*(1), 5–35.

Grimmelikhuijsen, S. G. and E. W. Welch (2012). Developing and testing a theoretical framework for computer-mediated transparency of local governments. *Public administration review 72*(4), 562–571.

Kohlschütter, C., P. Fankhauser, and W. Nejdl (2010). Boilerplate Detection using Shallow Text Features. In *Web Search and Data Mining*.

Marschall, M. and P. Shah (2013). Local elections in america project. *Center for Local Elections in American Politics. Kinder Institute for Urban Research, Rice University.(Database)*.

McNeal, R. S., C. J. Tolbert, K. Mossberger, and L. J. Dotterweich (2003). Innovating in digital government in the american states. *Social Science Quarterly 84*(1), 52–70.

McNutt, K. (2010). Virtual policy networks: Where all roads lead to rome. *Canadian Journal of Political Science/Revue canadienne de science politique 43*(4), 915–935.

Norris, P. (2003). Preaching to the Converted?: Pluralism, Participation and Party Websites. *Party Politics 9*(1), 21–45.

Osman, I. H., A. L. Anouze, Z. Irani, B. Al-Ayoubi, H. Lee, A. Balcı, T. D. Medeni, and V. Weerakkody (2014). Cobra framework to evaluate e-government services: A citizen-centric perspective. *Government Information Quarterly 31*(2), 243–256.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science 58*(4), 1064–1082.

Saperstein, D. (2003). Public accountability and faith-based organizations: A problem best avoided. *Harvard Law Review 116*(5), 1353–1396.

Sorens, J. (2018). The effects of housing supply restrictions on partisan geography. *Political Geography 66*, 44–56.

Stanyer, J. (2008). Elected representatives, online self-presentation and the personal vote: Party, personality and webstyles in the united states and united kingdom. *Information, Community & Society 11*(3), 414–432.

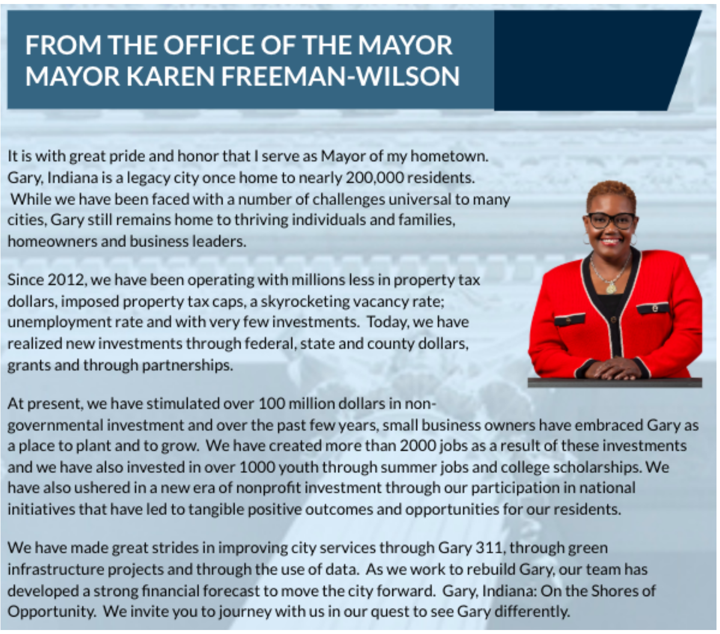
Sumner, J. L., E. M. Farris, and M. R. Holman. Crowdsourcing reliable local data. *Political Analysis*.

Thomas, J. C. and G. Streib (2003). The new face of government: citizen-initiated contacts in the era of e-government. *Journal of public administration research and theory 13*(1), 83–102.

Tolbert, C. J., K. Mossberger, and R. McNeal (2008). Institutions, policy innovation, and e-government in the american states. *Public administration review 68*(3), 549–563.

Valdez, D., A. C. Pickett, and P. Goodson (2018). Topic modeling: Latent semantic analysis for the social sciences. *Social Science Quarterly 99*(5), 1665–1679.

Wang, L., S. Bretschneider, and J. Gant (2005). Evaluating web-based e-government services with a citizen-centric approach. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pp. 129b–129b. Ieee.



FROM THE OFFICE OF THE MAYOR
MAYOR KAREN FREEMAN-WILSON

It is with great pride and honor that I serve as Mayor of my hometown. Gary, Indiana is a legacy city once home to nearly 200,000 residents. While we have been faced with a number of challenges universal to many cities, Gary still remains home to thriving individuals and families, homeowners and business leaders.

Since 2012, we have been operating with millions less in property tax dollars, imposed property tax caps, a skyrocketing vacancy rate; unemployment rate and with very few investments. Today, we have realized new investments through federal, state and county dollars, grants and through partnerships.

At present, we have stimulated over 100 million dollars in non-governmental investment and over the past few years, small business owners have embraced Gary as a place to plant and to grow. We have created more than 2000 jobs as a result of these investments and we have also invested in over 1000 youth through summer jobs and college scholarships. We have also ushered in a new era of nonprofit investment through our participation in national initiatives that have led to tangible positive outcomes and opportunities for our residents.

We have made great strides in improving city services through Gary 311, through green infrastructure projects and through the use of data. As we work to rebuild Gary, our team has developed a strong financial forecast to move the city forward. Gary, Indiana: On the Shores of Opportunity. We invite you to journey with us in our quest to see Gary differently.

Figure 1: Screenshot from the homepage at https://garyin.us/, accessed on 05/22/2019. Image depicts Democratic mayor of Gary, IN, Karen Freeman-Wilson.

#	Top Word 1	Top Word 2	Top Word 3	Top Word 4	Top Word 5	Top Word 6	Tokens assigned
49	artist	fun	music	beginner	player	prize	4565
46	chair	subcommittee	speaker	agenda	committee	commission	446
16	motion	second	adjourn	carry	unanimous	chairman	419
47	effluent	inf	eff	infiltration	discharge	sludge	751
21	everybody	think	something	thing	try	want	2609
2	influenza	infection	vaccine	patient	tuberculosis	hepatitis	2980
27	article	subsection	shall	franchisee	paragraph	meaning	658
30	subcontractor	bid	bidder	proposer	subcontract	bidding	512
12	craftsman	architecture	brick	distinctive	revival	storefront	1731
24	mail	fax	application	click	applicant	copy	367
34	playground	recreation	picnic	park	restroom	zoo	546
19	setback	variance	zoning	height	yard	accessory	453
26	mesa	canyon	via	odd	unidentified	paradise	1886
23	bag	recyclable	recyclables	reusable	vegetable	bait	2254
20	customer	renewable	efficiency	energy	saving	conservation	652
31	student	teacher	preschool	academic	kindergarten	youth	855
28	garland	assoc	association	firefighter	duke	xerox	480
50	trench	manhole	ductile	excavation	pipe	grout	1436
32	canceled	dwelling	suite	ave	tad	alteration	491
51	vent	combustible	flammable	egress	ceiling	extinguisher	1160
44	findings	tank	string	carcinogen	lust	sic	255
17	portfolio	micron	maturity	treasury	yield	investment	538
48	contributor	filer	officeholder	political	rouge	payee	293
5	draft	comment	review	revision	clarify	process	356
37	endorsed	endorse	rescue	assistant	analyst	technician	355
9	trust	revocable	planned	mfr	apportionment	exhibit	361
8	imp	assessor	taxpayer	petition	preliminary	determination	91
40	amt	invoice	acct	exp	unencumbered	encumbrance	116
57	councilman	introduced	alderman	whereas	resolved	councilwoman	615
11	obesity	sugary	epidemic	drink	calorie	sensible	96
15	credit	docket	app	post	download	month	61
3	wetland	specie	species	vernal	ecological	riparian	2293
29	margin	error	disability	speak	employed	language	180
43	medicare	payroll	blanket	contractual	undistributed	dept	322
42	incumbent	prep	batch	qualifier	analytical	examination	1091
55	taxable	deed	res	homestead	value	book	87
22	allocation	subtotal	admin	cost	yon	allocate	190
25	mitigation	impact	significant	adverse	environmental	measure	217
56	savings	neighborhood	village	excise	ltd	matrix	131
33	thence	east	south	corner	west	avenue	340
7	fugitive	bio	emission	coal	unmitigated	exhaust	773
18	perm	queue	delay	peak	adj	flt	187
54	license	licensee	citation	tow	fee	taxicab	710
6	race	householder	islander	census	occupied	female	160
60	bicycle	bike	pedestrian	route	sidewalk	bicyclist	561
14	accomplishment	grantee	narrative	outcome	grant	recipient	255
53	applied	col	dist	occupancy	monoxide	valuation	128
4	audit	auditor	procedure	timely	implemented	oversight	472
35	redemption	bond	increment	obligation	proceeds	lease	339
39	downtown	mixed	retail	waterfront	orient	density	419
10	grievance	deductible	coinsurance	dependent	employee	copay	583
38	para	persona	horas	bud	contracted	ante	1334
36	respondent	compare	figure	trend	appendix	satisfied	696
45	governmental	asset	actuarial	liability	financial	statement	235
41	complainant	allegation	defendant	offender	commander	complaint	1695
52	homeless	homelessness	affordable	supportive	housing	affordability	394
58	budget	revenue	adopted	balance	transfer	expenditure	176
13	initiative	outreach	strategy	leadership	engagement	focus	502
1	absent	preside	authorize	ordained	int	tag	377
59	burglary	robbery	theft	homicide	murder	gunshot	945

Table 1: Top words from a structural topic model with 60 topics and FREX scoring. Colors depict partisanship based on coefficient size. White cells are non-significant topics.