Content of Municipal Government Websites

Markus Neumann Bruce Desmarais Hanna Wallach

April 3, 2017

Abstract

We explore the effect of transitions of power in municipal governments on the content of their websites. We hypothesize that when party control changes, city administrators modify the contents of their websites in order to fit the agenda of the new incumbent. To test this theory, we apply an author-topic model [alternatively: structural topic model] to website data acquired from different points in time through the Wayback Machine.

1 Introduction

2 Background

Grimmelikhuijsen (2010) run an experiment in which citizens are exposed to randomly selected levels of information about local government council minutes. They find a negative relationship between the information level and perceptions of competence in the local government. This raises an interesting question regarding whether citizens or more likely to participate when they perceive competence or when they perceive incompetence.

Wang, Bretschneider, and Gant (2005) present a widely cited 'model' for evaluating the accessibility of information on government websites. This is an important paper with which we should be familiar at a very detailed level as we use archived web content to assess the volume/accessibility of information provided by local governments.

Osman, Anouze, Irani, Al-Ayoubi, Lee, Balcı, Medeni, and Weerakkody (2014) is less relevant, but they develop a multi-item measure to predict the level of citizen satisfaction with e-government services.

Grimmelikhuijsen and Welch (2012) conduct an enormously relevant study. Insofar as we analyze what predicts openness of government websites, we will be replicating and building upon this study. They focus on Dutch municipal websites, and their approach is fairly limited in scope and highly manual (which we can compliment). For example, one of the dependent variables "Decision-making transparency," is measured "using a discrete (1/0) indicator for whether the underlying principles or reasons for local air pollution policies were given on the Web site."

3 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 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

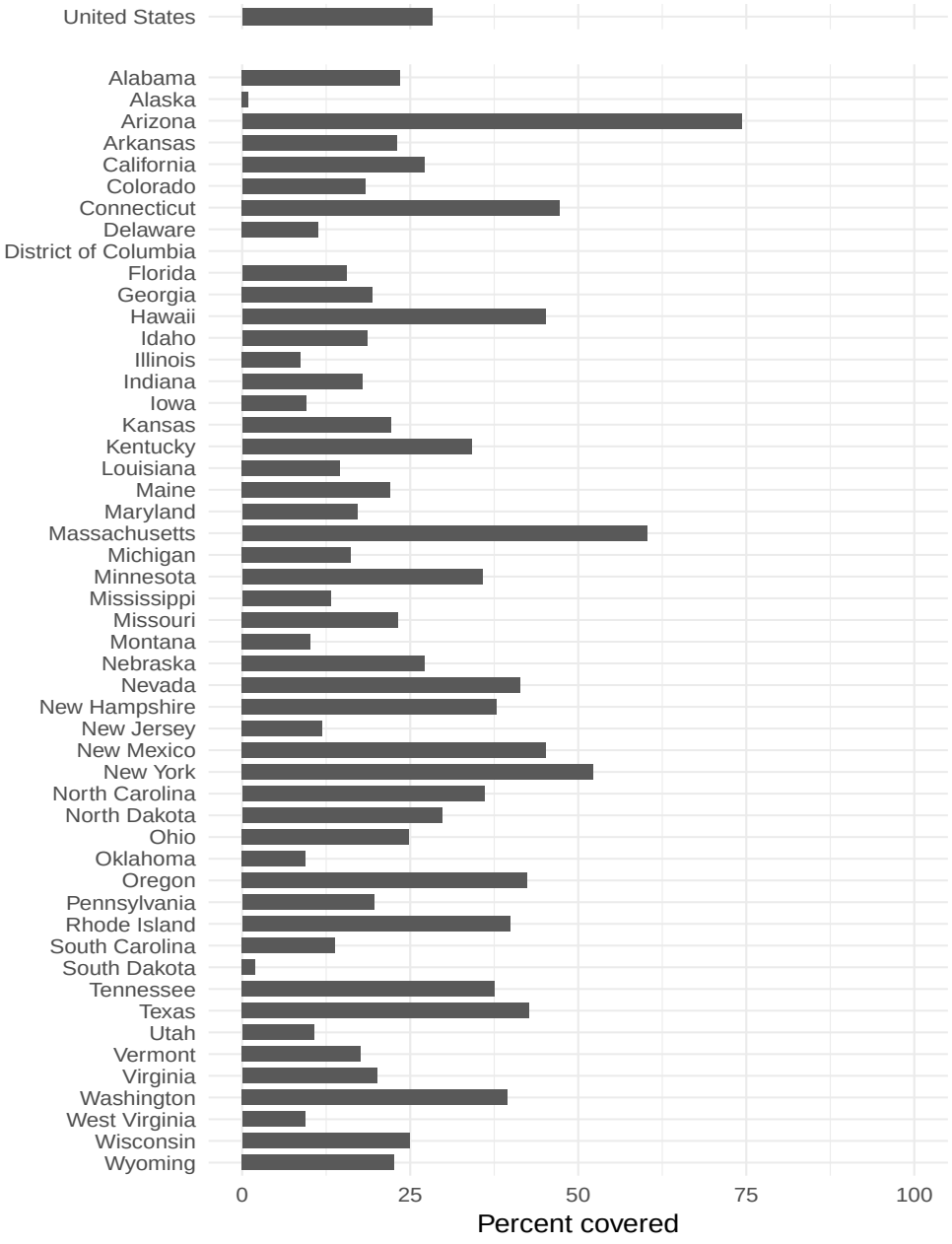
[1] Domains used for testing and internal programs are excluded.

[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

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

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



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.>

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

3.1 Indiana City Websites

It would be fine to focus on Indiana as a case. First, we need to answer some preliminary questions about the data.

1. For what percentage and number of IN cities can we find data from the WBM?

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

2. For how many election cycles can we find political leadership data for these matched cities?

3. In what number and percentage of cities is the local leadership majority Republican?

4. Relatedly, in a typical election cycle, for how many cities do we see a transition in party leadership (i.e., a shift from majority D (R) to majority (R) D).

1. 30 cities, with a combined population of 1,180,435. However, since only cities (as opposed to towns and villages) hold mayoral elections, only 16 of these, with a combined population of 1,094,383 can be matched to the election data.

2. 2015, 2011, 2007, 2003.

3. Of the 16 cities, 7 have Republican mayors after the 2015 elections.

4. In 6 cases, a shift of party control occurs, with 4 of these being Republican –> Democratic.

3.2 Research Design

1. Corpus:

   (a) Last snapshots before the election (November 3, 2015 in Indiana; tbd. in Louisiana (probably February))

   (b) First snapshot that is at least 2 months after the new government's inauguration (which is in January for Indiana, May for Louisiana)

2. Preprocessing:

   (a) restrict corpus to:

   i. documents belonging to cities in which a change of power occurred
   ii. documents that were added, deleted or changed between the two snapshots
   (b) words to lowercase
   (c) remove punctuation
   (d) stemming (Porter stemming algorithm?)
   (e) Remove stop words (regular list of stop words is enough, since we use an asymmetric prior)

3. Apply Grimmer's expressed agenda model to the corpus
   (a) Asymmetric prior
   (b) Each document can have only one topic (in contrast to the author-topic model)
   (c) Cities $i = 1, ..., n = 15$
   (d) Topic $k(k = 1, ..., K)$
   (e) Documents $j(j = 1, ..., D_i)$ from city i
   (f) Party covariate in the prior, where the deleted and unmodified documents are coded as from the first, and the added and modified documents from the second party

4. Results
   (a) Label topics using Grimmer's automatic cluster labeling method, based on most commonly used words in documents belonging to topic
   (b) Evaluate topics

Validation:

- Do the above for cities in which no change of power occurred.
- Check whether there is higher than average turnover around the new year by comparing changes to non-election years (and also Louisiana, where elections are later).
- Check how long documents stay on websites on average. Use websites with a lot of snapshots for this (these exist for both small and large cities).

Problem with using this model: Grimmer's expressed agenda model uses Senators as the actors. Senators is also who he is substantively interested in. For us, the equivalent to Senators is cities. However, we care about parties, not cities.

3.3 Topic modeling

Note: this chapter is mostly a wordy and less coherent version of the above.

We hypothesize that a change in leadership from one party to the other will lead to a change in website content because the two parties have different agendas. Democrats have a predilection towards policies that promote social and economic equality, whereas Republicans like to emphasize small government as well as law and order. Documents uploaded to city websites are expected to be a reflection of these preferences.

The Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is the most commonly used topic model. However, it is unable to account for the existence of two parties with very different policy agendas, translating to different preferred topics. There are two types of extensions to the LDA that fit our subject much better - the structural topic model, and the author-topic model.

The structural topic model, developed by Roberts, Stewart, and Airoldi (2016) allows researchers to model a corpus as a function of metadata associated with its documents. Specifically, topic prevalence (the proportion of a document made up by a topic) and topical content (the rate at which words figure into a topic) are contingent on a set of covariates. In our case, the two most important covariates are (1) city and (2) authoring party (operationalized by whether a document was present before a change of power, or introduced afterwards). Furthermore, the population size of the city should be a predictor for both the number and kind of problems it faces, which thus need to be addressed on its website. Furthermore, city size also serves as an estimate for the budget and technical capacities of its staff in charge of maintaining the website[6]. Further demographic as well as economic data might also be useful to differentiate cities from one another. If we model the differences between cities properly, we might not have to/should not include city as a (categorical) variable, because it would probably interfere with these more meaningful covariates.

The author-topic model (Rosen-Zvi, Griffiths, Steyvers, and Smyth 2004) would allow us to capture the fact that different authors have different topical preferences. Unfortunately, we have two types of 'authors' - cities, and parties. Given the largely divergent administrative needs of different types of cities, we would likely have to treat cities as the author. This would require us to capture the partisan authorship of different documents entirely on the basis of sub-sets of the website data - changed, added, or deleted documents. (Note: In the papers on author-topic models, the intention is often to analyze scientific articles. These articles are often co-authored. Would it be possible to have BOTH cities and parties as authors, so that a specific version of a website would then be 'co-authored' by its city and party?)

The critical element in this analysis is to accurately attribute authorship of documents to either party. Despite possible changes to websites due to a leadership transition, large parts of the content carry over. This means that unless the successor government decides to delete everything, some of the existing documents will be preserved, and in the model, also attributed to the new 'author'. But the reverse is not possible, because the predecessor government can't choose to retain documents

[6] Although this relationship is not exactly deterministic - when looking through .gov websites manually, I've noticed that a lot of websites of (presumably wealthy) towns of only a few thousand citizens often have extremely well-kept websites

from the future. *This is a very important point for municipal websites. We should investigate the possibility of modeling only the changes—documents that change, documents that are deleted, and documents that are added.*

Labeling newly added documents after a change of power is quite simple. As far as older documents are concerned, we would have to operate under the assumption that the incumbent didn't keep his or her successor's document's on the website for four years.[7] One problem here is the fact that the incumbent would have all the administrative topics assigned to them, simply because they have to have those on their website.

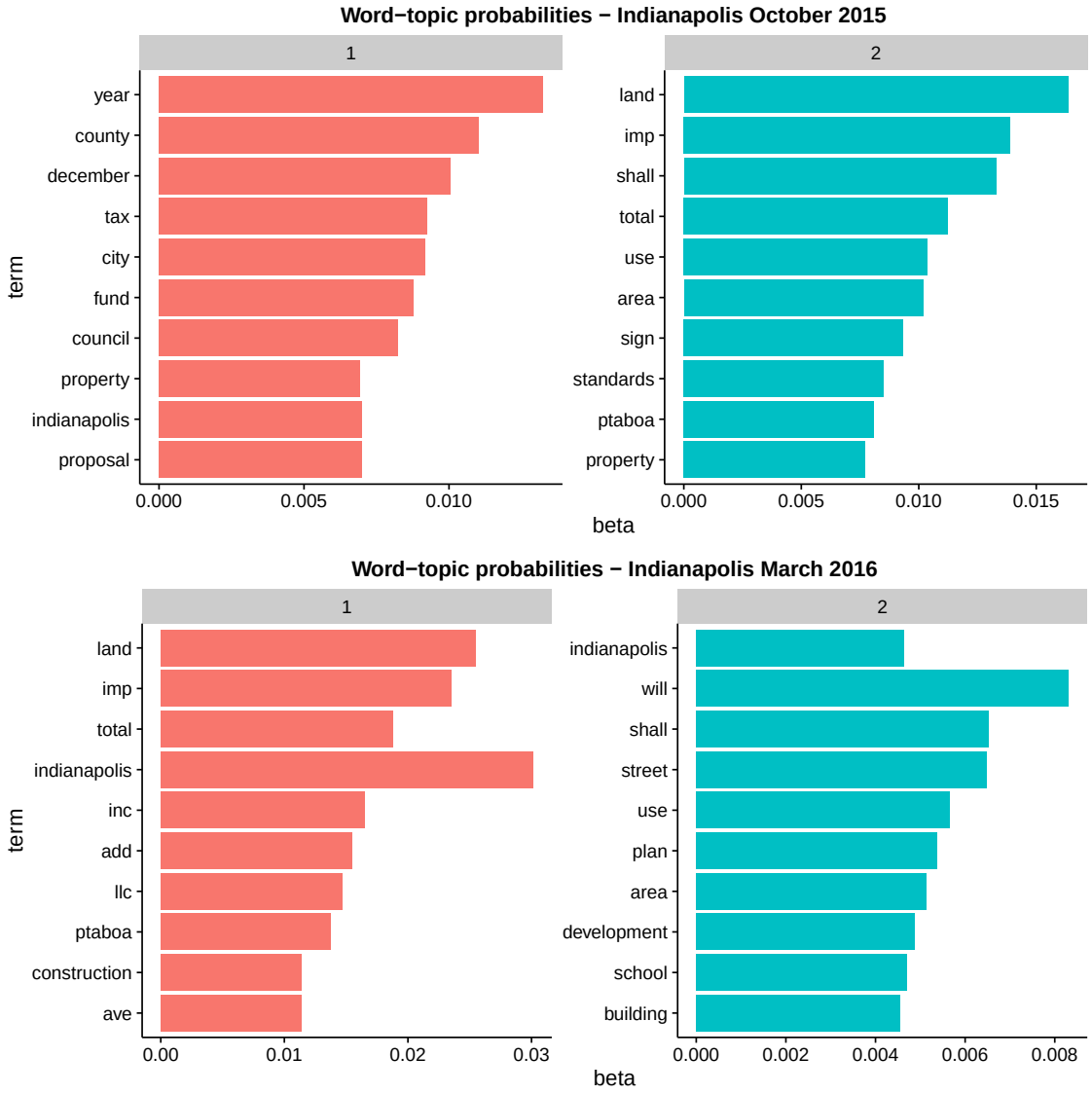
If we really do end up getting swamped with administrative terms in our topic models (and it does kind of look like that at the moment), we might be able to separate the signal from the noise by running a preparatory LDA once and using its results to create a new, corpus-specific list of stop words. After that, we run the actual model. This way, politically charged terms and topics, which likely are not as common, but present nevertheless, should be able to rise to the surface. It might be possible to refine this process by running an exploratory model on website data from cities in which party control never changes, and the incumbent always wins by large margins. 'Safe' cities like this should have fairly homogeneous populations, with little need for the incumbent to play politics on the municipal website. Hence, these websites should be filled with purely administrative content.

The use of asymmetric priors (Wallach, Mimno, and Mccallum 2009) over the document topic distribution - i.e. the assumption that some topics, such as administrative content, are inherently more common - may be a more elegant way of dealing with this issue.

Another intervening factor is that for cities in Indiana, mayoral terms begin in January. Since a lot of clerical and administrative tasks tend to be year-specific, work tends to pile up around the new year. Thus it is possible that a spike in newly added documents is not due to a change in party control, but owed to a seasonal increase in activity. We can test for this by comparing election years to non-election years. Furthermore, since in Louisiana, mayors take office in May, we have another point of comparison.

Furthermore, if we only investigate cities in which control of government changes from one party to another, we may overestimate its effect. Not only does a transition in party control occur, but the person in charge also changes. Parties are fairly homogeneous, so that two mayors from the same party may have very different policy preferences and managerial styles. To remedy this problem, we [could] utilize matching, pairing our cases with similar cities in which the incumbent does not run for re-election, but party control stays the same nevertheless.

[7] Probably a safe assumption. However, we could, and probably should test how long documents tend to stay on a city website. Simple descriptive statistics (for example density plots) on the length of existance should likely be sufficient. If we want to be really fancy about it, we could create a duration model, with document topics as features. This would allow us to measure whether some documents tend to remain longer based on their topic (i.e. fire regulations are probably going to stay up longer than notes on a specific council meeting).



References

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. ISSN 15324435. doi: 10.1162/jmlr.2003.3.4-5.993.

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

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

Ibrahim H Osman, Abdel Latef Anouze, Zahir Irani, Baydaa Al-Ayoubi, Habin Lee, Asım Balcı, Tunç D Medeni, and Vishanth Weerakkody. Cobra framework to evaluate e-government services: A citizen-centric perspective. *Government Information Quarterly*, 31(2):243–256, 2014.

Margaret E Roberts, Brandon M Stewart, and Edoardo M. Airoldi. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, pages 1–49, 2016. ISSN 0162-1459. doi: 10.1080/01621459.2016.1141684. URL http://www.tandfonline.com/doi/full/10.1080/01621459.2016.1141684.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, 2004. ISSN 01689002. doi: 10.1016/j.nima.2010.11.062. URL http://portal.acm.org/citation.cfm?id=1036902.

Hanna M Wallach, David Mimno, and Andrew Mccallum. Rethinking LDA: Why Priors Matter. *Advances in neural information processing systems*, 2009.

Lili Wang, Stuart Bretschneider, and Jon Gant. 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*, pages 129b–129b. Ieee, 2005.

	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

City	DemVotes	RepVotes	Winner	Change	Pop15	url
Attica		187	Republican	0	3117	https://attica-in.gov/
Connersville	1005	995	Democratic	1	13010	http://connersvillecommunity.com/
Frankfort		1748	Republican	0	16060	http://frankfort-in.gov/
Huntingburg	447	793	Republican	0	6035	http://www.huntingburg-in.gov/
Indianapolis	92830	56661	Democratic	1	862781	http://www.indy.gov
Lake Station	1483	227	Democratic	0	12054	http://www.lakestation-in.gov/
Linton	785	692	Democratic	0	5284	http://www.linton-in.gov/
Madison	1192	1915	Republican	0	12040	http://www.madison-in.gov/
Mitchell	229	495	Republican	1	4252	http://mitchell-in.com/
Monticello	0		Democratic	0	5322	http://www.monticelloin.gov/
North Vernon	679	697	Republican	1	6619	http://www.northvernon-in.gov/
Richmond	3421	2731	Democratic	0	35854	http://www.richmondindiana.gov/
Rockport	286	272	Democratic	1	2223	http://www.cityofrockport-in.gov/
South Bend	8515	2074	Democratic	0	101516	https://www.southbendin.gov/
Union City	338	440	Republican	0	3447	http://www.unioncity-in.gov/
Winchester	606	524	Democratic	1	4769	http://www.winchester-in.gov/

Table 4

City	snapshotPrior	electionDay	daysToElection
Attica	2015-08-01	2015-11-03	94
Cornersville	2015-10-18	2015-11-03	16
Frankfort	2015-10-18	2015-11-03	16
Huntingburg	2015-10-30	2015-11-03	4
Indianapolis	2015-10-16	2015-11-03	18
Lake Station	2015-10-29	2015-11-03	5
Linton	2015-08-01	2015-11-03	94
Madison	2015-09-05	2015-11-03	59
Mitchell		2015-11-03	
Monticello	2015-09-08	2015-11-03	56
North Vernon	2015-11-03	2015-11-03	0
Richmond	2015-10-31	2015-11-03	3
Rockport	2015-10-21	2015-11-03	13
South Bend	2015-11-01	2015-11-03	2
Union City	2015-09-25	2015-11-03	39
Winchester	2015-10-26	2015-11-03	8

Table 5: Closest snapshots to election day

File type	oct15	jan16	feb16	mar16
.pdf	179	183	111	197
.aspx	99	112	116	76
.jpg	22	24	17	12
.gif	3	5	1	5
.JPG	3	2	1	5
.png	3	7	2	1
.asmx	2	1	1	5
.PDF	1	1	1	2

Table 6: File types of Indy.gov at different points in time