Government Websites As Data: A methodological pipeline with application to the websites of municipalities in the United States*

Markus Neumann† Fridolin Linder‡ Bruce Desmarais§

October 10, 2019

Abstract

Objective: Existing social science research that uses government website content has relied on manual methods of content collection, limiting the scale and scope of projects. Our objectives and corresponding contributions in this research note are two-fold. First, we develop a methodological pipeline to gather, process, and analyze website content, along with an R package implementation of the pipeline. Second, we illustrate the pipeline through the collection and analysis of the websites of over two hundred municipal governments in the United States.

Methods: We propose and utilize a pipeline consisting of tools for web-scraping, text parsing and cleaning, and text analysis.

Results: Our pipeline is effective in extracting large-scale meaningful substantive text from government websites. We identify a strong relationship between the topical content on municipal websites and the partisanship of the mayor.

Conclusion: The pipeline we develop, associated software, and dataset of municipal websites, represent valuable research tools for social scientists.

*This work was supported by the National Science Foundation [1320219, 1637089, 1641047].

†Department of Political Science, The Pennsylvania State University, University Park, PA 16802, USA. Email: mvn5218@psu.edu. Corresponding author.

‡Department of Political Science, Social Media and Political Participation Lab, New York University, New York, NY 10012, USA. Email: fridolin.linder@nyu.edu

§Department of Political Science, The Pennsylvania State University, University Park, PA 16802, USA. Email: bdesmarais@psu.edu.