Website Data Codebook: 2016-2020

This codebook is a work in progress.

# Things to do:

* Create Counter Page for 2018, save script separate for 2019 and 2020

# Random Notes:

States not categorized: Delaware, Connecticut, Georgia, Illinois, indiana, Maryland, Minnesota, Washington,

# Directory Structure:

## Data\_Collection

* Scraping\_Websites
* Scraping\_BP
* Finding\_Websites

## Data\_Files

* Base\_Files\_Bank
  + State
    - Chamber
      * Candidate
        + HTMLs – Converted HTML’s from websites
        + PDFs – Scraped PDF images from websites
        + Raw\_Files – Untagged files
        + Sorted\_Files

Policy\_Pages – Pages with campaign policy statements

General\_Pages – Pages with content but not policy content

Junk\_Pages – Pages with nested HTML scripts or images/videos

* + - * + Text\_Files

Full\_Files – All pages with text converted into simple .txt showing only page text

Policy\_Statement\_List.txt – a single text file with split statements by <p> and <li> from all pages coded as policy pages

* Candidate\_Data
  + Website\_Data
    - A series of CSVs outlined within the Codebook below to denote different aspects of candidate information.
  + Associated\_Data
    - Folders with CSV’s associated with different projects and demographic variables.
* Sundry\_Data
  + unused or preliminary data, acquired and employed during data collection efforts.

## Data\_Processing:

# Codebooks:

## Candidate Identifiers

Candidates are tagged using a unique seven-digit identifier appended to name and state information. This is a consistent tag across all spread sheets, identified as *candidate\_id*.The final product looks like:

StatePostal-XXXXXXX\_LastName\_FirstInitial

### Candidate\_Name\_Information.csv

This contains a list of tags to identify individual candidates and can be used to connect different spreadsheets and different data sources.

*candidate\_ID*

The unique candidate identifier indicated above.

assigned\_ID

The unique candidate seven-digit identifier, composed of numbers and letters. Can also be found in *candidate\_ID.*

year

The corresponding year of candidate data. (Note: *candidate\_ID* and *assigned\_ID* extend across years in the data set.)

name

The name of the candidate.

first\_name

The candidates first name, derived from *name*.

last\_name

The candidates last name, derived from *name*.

bp\_url

The link to the candidates official Ballotpedia page for use with Ballotpedia.org data.

### Candidate\_District\_Information.csv

A file containing important information to identify the candidates state, chamber, and district. This can be joined to the Candidate\_Name\_Information.csv file to indicate the office a candidate ran for in a given year.

*candidate\_ID*

The unique candidate identifier indicated above.

assigned\_ID

The unique candidate seven-digit identifier, composed of numbers and letters. Can also be found in *candidate\_ID.*

year

The corresponding year of candidate data.

state

The full state name.

postal

The states postal abbreviation, capitalized.

FIPS

The states FIPS code, unpadded.

chamber

The chamber for which the candidate is running in that year.

long\_district\_FIPS

The long FIPS code for the candidate’s district, corresponding to the United States Census Bureau variable *GEO.id*.

long\_district\_name

The long district name, as identified by the United States Census Bureau variable *GEO.display-label*.

short\_FIPS

The short FIPS code for the candidate’s district, corresponding to the United States Census Bureau variable *GEO.id2*.

### Candidate\_Website\_Information.csv

A file containing information regarding the candidate’s website from the associated year including website url, a presence/absence variable, district identifier,party, and incumbency status.

*candidate\_id*

The unique candidate identifier indicated above.

assigned\_ID

The unique candidate seven-digit identifier, composed of numbers and letters. Can also be found in *candidate\_id.*

year

The corresponding year of candidate data.

website\_url

The URL for the identified candidate’s website in the associated year.

website\_binary

A binary variable, where 1 indicates the presence of a website.

party\_full

The candidate’s full party name.

incumbent\_status

A categorical variable indicating the candidate’s relative incumbency position within the race. I=incumbent candidate, C=challenger, O=open race

party\_short

A categorical variable indicating the candidate’s party, with 3rd parties lumped into a single group. D=Democrat, R=Republican, O=other

## Associated\_Data

Storage area for associated district or candidate data from that year.