United States Elections, Web Archives Data Package

Version information

Version 1.1 | Last updated 2024-10-31

Content Advisory

This data package includes an index of archived United States political campaign websites, social media, and other web content from federal, state, and U.S. territorial candidates 2000 - 2016. The material in this collection is political by nature. The indexes also include web content linked from those websites, which may contain embedded ads and user-posted comments. The content of these archived materials has not been reviewed or filtered.

I. Source Data or Collection Provenance

Which political office candidates are included?

This data package is comprised of index files from the United States Elections Web Archive, which includes campaign websites documenting presidential, congressional, and gubernatorial elections that were archived weekly during general election seasons since 2000. The data package currently includes years 2000 – 2016.

Prior to election 2020, the sites archived in this collection often include web-harvested social media content, in order to provide a fuller representation of how candidates presented themselves via the Internet to the electorate. However, due to the nature of social media platforms, success rates for capturing social media content have varied.

Note that not all types of candidates are included for all years.

	Presidential candidates	Congressional candidates	Gubernatorial candidates	Social media of federal candidates
2000	X			
2002	X	X	X	X

2004	X	X	X	X
2006	X	X	X	X
2008	X	X	X	X
2010	X	X	X	X
2012	X	X	X	X
2014	X	X	X	X
2016	X	X	X	X

Which political campaign websites are included?

General criteria for including candidates' websites:

- The election occurred during a general election year (even-numbered years) and was not an off-year election (e.g., special elections, and gubernatorial elections for Louisiana, Mississippi, New Jersey and Virginia).
- Presidential candidates: based on campaign committee website URL provided via FEC campaign data and supplemented by staff research as needed
- House candidates: based on campaign committee website URL provided via FEC campaign data and supplemented by staff research as needed
- Senate candidates: based on campaign committee website URL provided via FEC campaign data and supplemented by staff research as needed
- Social media sites of presidential, house, and senate candidates, as identified by information on the campaign website or online searches
- Gubernatorial candidates: online staff research

Presidential, House, and Senate candidates

In 2000 and 2002, candidates and candidate website data was gathered through online research by Library of Congress staff and contractors. Beginning in 2004, the list of campaign websites to include for presidential and congressional candidates has been pulled from Federal Elections Commission (FEC) campaign data and supplemented by research from Library of Congress subject specialists. Beginning in 2006, the FEC data has been automatically pulled from the API OpenFEC. In 2004, the data was manually retrieved from the FEC website.

The FEC data is submitted by candidates, including a website URL of their primary campaign committee. If a candidate's campaign committee does not supply a website, and Library staff are not able to locate a website, the web archive may not include a website for that candidate in that election year. Library of Congress subject specialists review and enhance the FEC data based on online searches.

The Library of Congress attempts to capture websites of all congressional candidates who win their party's primary and appear on the general election ballot, regardless of party affiliation or "final election results". For some of the early years of the collection, candidates may also be included who made primary ballots but not the final general election ballots. For presidential candidates, the Library attempts to collect the websites of all major candidates who announce campaigns. The Library also collects notable third party and independent candidates with general election ballot access to at least 50 electoral votes.

Gubernatorial candidates

The list of websites to include for gubernatorial candidates is manually compiled by Library of Congress staff based on online searches. An attempt is made to capture the websites of all gubernatorial candidates who have won their party's primary and appear on their state ballot, regardless of party affiliation or winning candidate. For some of the early years of the collection, candidates may also be included who made primary ballots but not the final general election ballots.

How about social media accounts?

From 2002-2020, social media channels were sometimes captured, either as stand-alone websites or they were scoped in when they were linked from the included campaign websites. Using either method, these captures may not be a complete representation of all social media content produced by included candidates. Note that this data package currently contains data only through 2016.

When social media sites were archived using the second method, they were collected as "scopes", meaning that the web archive crawler is instructed to collect a candidate's primary campaign website and the social media "scope" if it is linked from the primary campaign website. For example, Barack Obama's 2012 primary campaign website is , and it had a scope added for twitter.com/BarackObama. As the crawler collected webpages, any links it came across to URLs beginning with twitter.com/BarackObama it would collect as well. In this data package's metadata.csv file, these social media sites appear in the `website_scopes` column.

In some cases, particularly when a traditional campaign website was not identified, social media channels may have been collected as target websites, called "seeds." In other cases, as social media became more challenging to archive, they were collected as seeds to try to ensure more comprehensive captures. For example, in this data package is a primary seed for Illinois Kevin W. Yoder's campaign for Illinois' 3rd House district from 2010 through 2016, along with .

When are websites collected?

Although there was some variation in earlier years, for presidential elections, the team has typically begins capturing websites in the lead up to primaries. For gubernatorial and congressional elections, the team generally begins capturing websites in general election years when candidates have been selected for the final ballot, after a state's primary election results are announced. The frequency with which these websites are crawled can vary, but is usually once a week, and the content becomes publicly available on the United States Elections Web Archive collection after an embargo period of no less than one year. Off-year elections (e.g., special elections, and gubernatorial elections for Louisiana, Mississippi, New Jersey and Virginia) are not included in this data package. This is due to the way that CDX indexes are organized and intermingled with other content.

Is there election content not in the indexes?

Yes. Election content collected outside the general election cycle, either in off-year elections or collected early in regular election years before collection for the United States Election Web Archive formally begins around May to June, is not typically included in the CDX index files for this data package. Those records will be intermingled within the CDX files for regular crawls across all collections, which are not in scope for this data package

Also, if a site owner chooses to exclude their site from the collection or submits a takedown request, after Library review those sites may either be unavailable or only available onsite. For unavailable content, this content will not appear on the metadata.csv file, may be absent from the CDX files, and the web documents will not be found in the archive.

Is there other content in the index?

Yes, there is a significant amount of "noise" in this index. This additional content comes from several sources, some of which may be of interest to your research project and others not.

1. Non-candidate websites

In the early years of the United States Elections Web Archive collections, the Library of Congress included websites of political parties, government, advocacy groups, bloggers, and other individuals and groups producing content relevant to the election. These sites have generally been moved into the Public Policy Topics Web Archive or into the general web archives. However, because these sites were originally collected with candidate websites, they remain in the CDX index files in this data package. This data package reflects the scope of the United States Elections Web Archive at the time of its collection each general election year. Because this scope has evolved, the scope of the CDX index files also varies over time.

Additionally, candidates have not always used personal campaign websites. Some candidates websites are simply pages on their party's organization website. For example, the 2004 campaign website for Steven Larchuk, the American Healthcare Party candidate for

Pennsylvania's 4th district, used the website http://healthcareparty.sitebuilder.completecampaigns.com/.

2. Websites and files linked from candidate websites

The web archive is built using automated tools, and the technical configuration of these tools has an important impact on the scope of the collection. When a website is selected for inclusion, it is archived using a web crawler. The crawler starts with a "seed URL"— for instance a homepage—and the crawler follows the links it finds, preserving content as it goes. These links include HTML, CSS, JavaScript, images, PDFs, audio and video files, and more. Scoping instructions are added to allow or restrict the crawler's ability to collect linked files hosted on third party sites or on other subdomains from the same organization. The resulting web archive contains a broad diversity of file formats, from a range of web domains not necessarily limited to those originally selected for inclusion. More information about the crawling process can be found at https://www.loc.gov/programs/web-archiving/about-this-program/frequently-asked-questions/.

See also

- Five Questions for Will Elsbury, Project Leader for the Election 2014 Web Archive (2014-10-16 blog post)

A. Historical background of source material

Campaign committees of U.S. federal and gubernatorial candidates commonly create campaign websites and social media accounts during election races. The Library's Web Archiving Program began the United States Election Web Archive in 2000, when the concept of a campaign website was still relatively new, and only a fraction of Americans reported getting election information online.

Original format

The web archives are made up of a range of individual file formats including HTML, CSS, JavaScript, image, PDF, audio, video, and other types of files.

B. Acquisition and Access

Library of Congress reading room and contact

The candidates and their website titles can be searched via the United States Elections Web Archive collection. Metadata about each candidate, such as political party and election location, can be searched, but the full text of the websites themselves cannot be searched. Individual URLs from the collection can be browsed by adding the URL to the end of **https://webarchive.loc.gov/all/*** as in https://webarchive.loc.gov/all/*/http://johnlewisforcongress.com/

Metadata type

The CDX index files contain minimal metadata about each document URL archived in this collection. A profile of the CDX format and its metadata is at https://www.loc.gov/preservation/digital/formats/fdd/fdd000590.shtml. This collection specifically uses the 2015 version of the CDX specification, which includes 11 fields.

The metadata.csv file uses a locally defined set of columns.

For CDX and metadata.csv fields, see IV. Dataset field descriptions.

Scale of description

The CDX files represent archived websites at a very granular level. Specifically, each line in a CDX file describes one file that has been archived, including images, javascript, css stylesheets, PDFs, HTML, etc. (e.g., https://warnockforgeorgia.com/about/ or https://warnockforgeorgia.com/wp-includes/js/jquery/jquery.min.js?ver=3.7.1).

C. Digitization

No digitization has been performed on this collection. All contents are received-digital.

II. About the exploratory data package

This data package includes 396,117 pointer index files, totaling 223.9 GB, from 2000 - 2016. These index pointer files can be used to download all archived files from the collection, for included years. The pointer files are in the CDX format and can be downloaded in bulk or subdivided by election year.

The data package also includes a CSV metadata file listing all political candidate campaign websites. Because the CDX pointer index files include a broad and varying scope of materials (see Is there other content in the index?), this CSV file can be used to filter down to a subset of domains within the CDX files, exclusive to political candidate websites. See the Python notebook demonstration code associated with this data package for more information about how to use this metadata file to filter the CDX indexes.

See also:

- Candidates, Campaigns, and CDX Files: A New United States Elections Web Archive Dataset (2024-04-13 blog post)

- The Library of Congress Web Archives: Dipping a Toe in a Lake of Data (2019-01-09 blog post)

What's included?

The data package contains:

- README.md - technical overview of how all assessment datasets were created and the context in which they were created
- 396,117 CDX index files (223.9 GB) - compressed with gzip and organized by election year
- metadata.csv (8.4 MB) - list of all websites and scopes websites in the United States Elections Web Archive, useful for filtering the CDX index lists.

Composition

	CDX file count	Total CDX size, gzipped (GB)	CDX URL base
2000	30,521	1.8	https://data.labs.loc.gov/us-elections/2000/*
2002	9,484	4.9	https://data.labs.loc.gov/us-elections/2002/*
2004	48,451	10.2	https://data.labs.loc.gov/us-elections/2004/*
2006	95,797	29.0	https://data.labs.loc.gov/us-elections/2006/*
2008	141,288	38.4	https://data.labs.loc.gov/us-elections/2008/*
2010	8,993	31.8	https://data.labs.loc.gov/us-elections/2010/*
2012	16,211	29.1	https://data.labs.loc.gov/us-elections/2012/*
2014	21,541	19.2	https://data.labs.loc.gov/us-elections/2014/*
2016	50,831	59.5	https://data.labs.loc.gov/us-elections/2016/*

The CDX files are space-delimited text files that can be read by a plain text editor once unzipped. They are structured somewhat similarly to a tab-delimited text file or a CSV, with some important distinctions.

The first five lines of a CDX might look like:

```
CDX N b a m s k r M S V g

com,buzzfeed,s)/static/campaign_images/2008/11/5/13/5a50b25ce025246c6f9f0e0aa
```

```
org,iste)/content/navigationmenu/membership/sigs/sigcs_computer_science_/jctj

org,norml)/sendpage2.cfm?wtm_reference=<http://www.norml.org/index.cfm?group_

com,keystonepolitics)/comment/reply/17022/59971 20081108010600 <http://www.ke
```

The first line of each CDX file will typically be `CDX N b a m s k r M S V g`. The first character (a space) indicates which character serves as a delimiter separating fields on each line. The following "CDX" declares that this is a CDX file. The eleven letters that follow ("N" through "g") are codes for various fields in the data. See IV. Dataset field descriptions > CDX files for details about these field definitions.

Each of the following lines in a CDX is a web file that has been archived, such as an html, css, javascript, image, PDF, audio file, etc. Note that even small variations in the URLs to these files will be represented as additional lines in the CDX. For example, for the purposes of the CDX www.mittromney.com/ and mittromney.com/ would be considered distinct file URLs and would be represented on separate lines.

For more information about the CDX format, see the 2015 version of the CDX specification.

For information about what type of record each line represents, see ***Scale of description.***

For information about the data fields ("N" through "g"), see ***IV. Dataset field descriptions***.

For information about how to use CDX files as pointers to download archived websites, see the demonstration Python notebook attached to this data package.

Data inconsistencies

Please also note that the data included are the raw CDX files, which may be malformatted or contain errors that do not conform to the CDX specification. Please see the demonstration Python notebook attached to this data package for more guidance on cleaning the data.

Potential risks to people, communities, and organizations and strategies for risk mitigation

The CDX files include minimal metadata about public websites archived by Library of Congress. The content of this collection becomes publicly available on the United States Elections Web Archive collection after an embargo period of no less than one year.

Computational readiness and possible uses

From this dataset, a user can generate pointers to download all archived web content from the United States Elections Web Archive. This dataset does not include the collection content itself, only pointers.

The CDX files included are the raw files from storage, and data cleaning will be necessary for most uses. Please also note that the data included are the raw CDX files, which may contain errors that do not conform to the CDX specification. Please see the demonstration Python notebook attached to this data package for more guidance on cleaning the data.

Because the CDX pointer index files include a broad and varying scope of materials (see Is there other content in the index?), researchers may opt to use the metadata.csv file to filter down to only political candidate websites within the CDX file. See the Python notebook demonstration code associated with this data package for more information about how to use this metadata file to filter the CDX indexes.

III. How was it created?

The web archive CDX indexes are created as part of a routine process to provide access to archived web objects in the Library of Congress's online web archive access system, unrelated to the creation of this data package.

Compilation Methods

The CDX files in this data package were gathered from their internal storage locations and collated into a public AWS S3 bucket for download, by staff of the Library of Congress Web Archiving Program. This process does not use any public APIs, and the CDX files are not publicly available.

Preprocessing steps

The contents of the CDX files have not been altered. No CDX files were filtered out of the data package; all CDX files found on storage for a given election year are included in this data package.

Files were organized by election year, which does not reflect the original directory structure of the files. The 2022 files were compressed using gzip, because the original files were not compressed unlike other years.

IV. Dataset field descriptions

metadata.csv

Each row in the metadata.csv file is a candidate website collected within a specific election year. For example, the spreadsheet has five rows for Mitt Romney:

1. 2002 Massachusetts Governor campaign: http://www.romney2002.com/
2. 2008 Presidential campaign: http://mittromney.com/
3. 2008 Presidential campaign: http://www.mittromney.com/
4. 2012 Presidential campaign: http://mittromney.com/
5. 2012 Presidential campaign: http://www.mittromney.com/

The metadata.csv file uses a local set of columns, as described in the table below. The following terms are used in the table:

- **Item –** a loc.gov record, such as https://www.loc.gov/item/lcwaN0000512/, which represents one person who has been a U.S. political candidate.
- **Website –** the base URL of a website associated with an item (a candidate) for one or more election years. In common web archiving terminology, this is equivalent to a "seed".

Field	Level	Datatype	Definition	Metadata Source
item_id	Item	String	URL to the item record in loc.gov. Each item corresponds to one candidate. E.g., http://www.loc.gov/item/lcwaN0000528	loc.gov MODS record
item_title	Item	String	Title of the item as it appears in loc.gov. E.g., "Official Campaign Web Site - Joe Donnelly"	loc.gov MODS record
website_url	Website	String	URL to the archived website (aka, the "seed"). E.g., http://www.donnellyforuscongress.com/	loc.gov MODS record
website_id	Website	String	5-digit unique ID for the seed in the Library of Congress's backend systems. E.g., 00852	loc.gov MODS record

website_scopes	Website	List	Additional websites that the web archive crawler was instructed to also capture if the seed website linked to them. NOTE: this is a list of all scopes ever added to this record across election years.	loc.gov MODS record
collection	Record	String	The collection (election year). E.g., United States Elections 2008	loc.gov MODS record
website_elections	Website	List	A list of campaigns that the candidate ran in for a given year. Usually, this is a list of only one election. However, if a candidate dropped out of one campaign and started another in the same election year, this list will have more than one item.	loc.gov MODS record
website_parties	Website	List	List of parties, for each item in the list of campaigns. The ordering of this field aligns with the election field.	loc.gov MODS record
website_places	Website	List	List of campaign locations, for each item in the list of campaigns. The ordering of this field aligns with the election field.	loc.gov MODS record

website_districts	Website	List	List of campaign House districts, for each item in the list of campaigns. The ordering of this field aligns with the election field. For non-House campaigns, the None is used.	loc.gov MODS record
website_thumbnail	Website	String	URL to the see thumbnail, as seen on the item page in loc.gov	loc.gov MODS record
website*start*date	Website	Integer	The first meaningful (non-404) capture of this seed.	loc.gov MODS record
website*end*date	Website	Integer	The last meaningful (non-404) capture of this seed.	loc.gov MODS record
item*all*years	Item	List	All election years for which this item (candidate) is included in the United States Elections Web Archive.	loc.gov MODS record
website*all*years	Website	List	All election years for which this particular website is included in the United States Elections Web Archive.	loc.gov MODS record
mods_url	Item	String	URL to the MODS metadata record linked from the candidate's loc.gov item page. This file is the source of all other metadata fields.	loc.gov item record

access_condition	Item	String	Conditions of access. "None" if the materials are publicly available online, or "Access restricted to on-site users" if the materials can only be accessed onsite from Library of Congress's DC-area locations.	loc.gov item record

CDX files

This section lists and describes each of the fields included in the United States Elections Web Archive CDX indexes as they align with 2015 version of the CDX specification. The CDX indexes contain 11 fields (listed in the first line of each CDX file), with the corresponding information for each field as follows. "Metadata Source" indicates the source of the CDX field.

Please also note that the data included are the raw CDX files, which may contain errors that do not conform to the CDX specification. Please see the demonstration Python notebook attached to this data package for more guidance on cleaning the data.

Field	Name	Datatype	Definition	Metadata Source
N	urlkey	String	The URL of the captured web object, without the protocol (http://) or the leading www and in SURT format (http://crawler.archive.org/articles/user_manual/glossary.ht	WARC (Web ARChive) or ARC preservation archival file
b	timestamp	Integer	timestamp in the form YYYYMMDDhhmmss. The time represents the point at which the web object was captured, measured in GMT. This field and the `original` field can be combined to generate a download URL for the archived web document.	WARC (Web ARChive) or ARC preservation archival file

a	original	String	The URL of the captured web object, including the protocol (`http://` or `https://`) and the leading www, if applicable. This field and the `timestamp` field can be combined to generate a download URL for the archived web document.	WARC (Web ARChive) or ARC preservation archival file
m	mimetype	String	The media type (https://www.iana.org/assignments/media-types/media-type as provided by the archived website's server. Note that web servers do not always correctly identify the media type of files.	Original web server, via the WARC (Web ARChive) or ARC preservation archival file
s	statuscode	Integer or "-"	The HTTP response code received from the server at the time of capture, e.g., 200, 404, or "-" for lines that represent requests or metadata.	WARC (Web ARChive) or ARC preservation archival file
k	digest	String	A unique, cryptographic hash of the web object's payload at the time of the crawl. This provides a distinct fingerprint for the object; it is a Base32 encoded SHA-1 hash.	WARC (Web ARChive) or ARC preservation archival file
r	redirect	String	This is usually "-" or blank.	WARC (Web ARChive) or ARC preservation archival file
M	metatags	String	HTML meta tags, from the header of HTML files. This is usually "-" or blank.	HTML website file, via the WARC (Web ARChive) or ARC preservation archival file

S	file_size	Integer	The size of the web object, in bytes	WARC (Web ARChive) or ARC preservation archival file
V	offset	Integer	The location of the corresponding record in the original uncompressed Web Archive file (WARC or ARC), which stores the full archived object and is not publicly accessible.	WARC (Web ARChive) or ARC preservation archival file
g	WARC filename	String	Name of the compressed Web Archive (WARC or ARC) file, which stores the full archived object and is not publicly accessible.	WARC (Web ARChive) or ARC preservation archival file

V. Rights Statement

Rights for this data package

The README, CDX files, and metadata.csv contained within this data package have no known copyright restrictions and are free to use and reuse.

Rights for the source United States Election Web Archive collection:

See the full Rights & Access statement for the United States Election Web Archive at https://www.loc.gov/collections/united-states-elections-web-archive/about-this-collection/rights-and-acces

VI. Creator and contributor information

Creators: Rachel Trent, Chase Dooley

Contributors: Grace Bicho, Lauren Baker, Abbie Grotke, Tracee Haupt, Amanda Lehman, Ken Drexler

VII. Feedback

Please direct all questions and comments to the Web Archiving Program at webcapture@loc.gov. We welcome you to share with us information or links to your research using this data package!