

Resumes of Congress

Data Scrape and Validation

March 2024

Table of Contents

- Introduction
- Objectives and Scope
- PDF Data Extraction Methodology
- PDF Data Extraction
- Data Prep and Validation
- Findings and Conclusion

Introduction

Since 1947, Congress has published a Resume of Activity at the end of each session. This document is part of the Congressional Record and includes:

- Length of session
- Legislative measures introduced, reported, and passed
- Votes
- Disposition of executive nominations (civilian and military)

Objectives

Scope:
98th through 117th
Congresses

**PDF Data
Extraction
Evaluation**

Compare Python PDF libraries for accuracy and ease of use

**PDF Data
Extraction**

Create a dataset from the Resumes of Congressional Activity

**Prep and
Validation**

Tidy, validate, and publish the final dataset for use in future analysis projects

PDF Data Extraction Methodology

Project Objective #1

Compare Python PDF libraries for accuracy and ease of use

Formats

Resumes for the 98th through the 117th Congresses use two distinct formats.

Resumes before 1997 are available as scanned images.

98th – 105th Congresses:
1 file, 2 pages

Page 1:
**General
Activity**

Session 1	Session 2

Page 2:
**Confirmation
Data**

Session 1	Session 2

105th – 117th Congresses:
2 files, 1 page each

File 1:
Session 1






General Activity	Confirmation

File 2:
Session 2

General Activity	Confirmation

Methods

Each library was tested using the Final Resume of Congressional Activity for the 98th Congress.

Tool		Result *
pypdf 4.1.0		Reads text but does not recognize tables
pdfminer 20191125		Reads text by column from top to bottom, separating labels from values
tabula-py 2.9.0		Recognizes tables and reads text successfully into dataframes
camelot-py 0.9.0		Results are similar to tabula, but output is not as clean
pdfplumber 0.11.0		Reads text but does not recognize tables; splits words

* Code used for testing is available on [GitHub](#).

PDF Data Extraction

Project Objective #2

Create a dataset from the Resumes of Congressional Activity published annually in PDF format

Challenges

Intent:
Convert Resumes
from PDF format to
Excel using tabula.py

Inconsistent
File Formats

The object structure of the PDF files varies significantly from year to year.

Automation is
Hard

tabula.py did not produce consistent results when run against a series of files.

Challenge

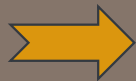
Balance time and quality by combining manual and automated processes

Extract Methodology

Due to inconsistent results using the tabula.py library, a hybrid extraction methodology was used:



OCR in Acrobat
Pro, if needed



Extract to CSV
using tabula app



Convert to Excel



Correct text
errors in Excel



Run macros and
format

Legislative Activity Data Cleanup Macro

Purpose:

Automate the most common formatting errors in the tabula application extract

Scope:

- Combine headings that have been split across multiple columns and/or rows
- Remove special characters from numeric columns
- Add data to indicate the Congress, Session, Start Date, and End Date
- Resize columns and align text

Confirmation Data Cleanup Macro

Purpose:

Automate the most common formatting errors in the tabula application extract

Scope:

- Delete empty cells
- Combine headings that have been split across multiple columns and/or rows
- Extract values that have been embedded into section headings
- Remove trailing and leading periods
- Add data to indicate the Congress, Session, Start Date, and End Date
- Resize columns and align text

Data Prep and Validation

Project Objective #3

Tidy, validate, and publish the final dataset for use in future analysis projects

Data Preparation

Data preparation focused on cleaning up inconsistencies in how data was labeled from year to year, including:

- Separating data from labels, especially in the nominations data
- Standardizing data labels/column headings
- Splitting new nominations from carryover nominations for consistency
- Reformatting date and time columns

Data Preparation

- Merging of similar types of data where categorization and labeling was inconsistent from year to year:
 - Civilian data
 - Failed dispositions
 - Returned Dispositions
 - Recess Reappointments

Data Validation

Verify Data Scrape / OCR

- Verify data is present for all congressional sessions in scope
- Verify data is present for both houses of Congress
- Verify detail lines roll up properly and match category totals

Verify Data Integrity

- Review and update missing data
- Verify session dates against Congress.gov
- Verify days and time in session are reasonable
- Check for data outliers

Findings

The data published in the Resumes of Congressional Activity are prone to errors. Example:

In the second session of the 112th Congress, the total Measures Introduced in the Senate is 2,447. However, the total of Bill and Resolutions Introduced is 2,448.

A complete list of data errors is documented in the Data Validation Issues document available in the GitHub repository for this project.

Findings

A total of 56 data errors were documented.

Data errors were identified in 19/40 data files, or 48%.

Data files with errors averaged 3 errors per file.

Conclusions

This project resulted in the creation of two data files– one for Confirmation data and one for Legislative data.

The data does contain some errors, which have been documented for future reference.

Possible next steps:

- Conduct further research to resolve data errors
- Expand the dataset to include more sessions of Congress
- Explore the dataset for trends

Thank You

Tami McManus

https://github.com/tamimcm416/congressional_activity

<https://www.linkedin.com/in/tami-mcmanus/>

Appendix A – Source Data

“Resume of Congressional Activity (1947 to Present)”, United States Senate, [PDF], <https://www.senate.gov/legislative/ResumesofCongressionalActivity1947present.htm>. Accessed Nov 23, 2023.

“Days of Past Session”, United States Congress, <https://www.congress.gov/past-days-in-session>. Accessed Mar 24, 2024.

Appendix B – Tools Used

Adobe Acrobat Pro, Version 2024.001.20629

JupyterLab, Version 4.1.2

Microsoft 365, Microsoft Excel, Version 2403

Microsoft Visual Basic for Applications, Version 7.1

Python, 3.12.2

tabula, Version 1.2.1

Appendix C – Image Attribution

Acrobat Pro thumbnail:

<https://uxwing.com/acrobat-pro-icon/>

tabula thumbnail:

<https://github.com/tabulapdf>

Microsoft Excel thumbnail:

https://commons.wikimedia.org/wiki/File:Microsoft_Office_Excel_%282019%E2%80%93present%29.svg

Keyboard thumbnail:

<https://www.iconpacks.net/free-icon/keyboard-1385.html>

Visual Basic for Applications thumbnail:

https://www.iconfinder.com/icons/4195441/macros_spreadsheet_vba_icon