Hi Guy:

Thanks for taking this on. There’s a lot of potential in the project given that data are now more easily available than when we started (we actually had to scan and OCR pages from the house and senate journals). I’m sure you’ll have questions once you get into it and I’m ready to help.

An undergraduate, Rohnin Randles, wrote the scripts and excellent descriptions for collecting most of the bill variables.

<https://github.com/rwrandles/cbp>

Unfortunately, Rohnin ran out of time (he just started in the Princeton Ph.D. program), so Ph.D. student Zhaowen Guo stepped in to finish things off.

* Bill\_merged.R – This script applies sponsor variables collected for the previous Congress to the next Congress. This may be helpful because a lot of the information for incumbents transfers

These three scripts are used to create variables indicating whether the sponsor served on a committee to which the bill was referred, in addition to other variables (such as a list of the committees of referral etc)

* Legislator\_committee - collects information about bill sponsor committee positions using the bioguide id as the identifier
* Referral\_committee – collects information about bill and resolution committee referrals
* Merge\_sponsor\_committee.R – uses the information from the two previous scripts to create several new variables – ComC, ComR, mref, chref, rankref (see variables codebook)

**Additional information**

The data currently go through May 2020. Updating to the end of the current Congress should be fairly straightforward because the member variables for the 116th Congress are already in place.

My advice would be to update annually or bi-annually, starting this December or January.

There are several components. The first, obviously, are the topic codes. The second is the information about the bills themselves and the third is the sponsors.

**Topic codes**: For the most recent dataset on this page (<http://congressionalbills.org/download.html>

) , we used a random forests supervised algorithm to label the m. We first predicted the major topic code, and then predicted the subtopic for the bills within each major topic. This process seemed to produce very accurate results at both the major and subtopic levels. The other datasets on that page were manually labeled and we used them for training of course. The topic codes are based on the bill titles only (not the summaries or text).

We only assign topics to HRs and S (not resolutions, although one can argue that joint resolutions should also be coded because they have the force of law). The reason we haven’t done everything is because it would be overwhelming for manual annotation. But certainly worth considering automated annotation for all of them.

Scripts for doing this can be found on Rohnin’s github page (see below).

**Bill and Sponsor Variables:** The variables codebook is here: (<http://congressionalbills.org/codebooks.html>). The process of creating each of these variables has always been a challenge because they originate from different sources. The good news is that this has gotten much easier thanks to the <https://github.com/unitedstates/congress> website and the related API <https://projects.propublica.org/api-docs/congress-api/> . The API is easier but it is not as complete as the scripts available from the github repository so we use both. Nearly all of the variables can be scraped from this single resource, or converted from information available there. Yeah!

However, we are relying on someone else’s scripts, which may contain errors or have fields that do not perfectly correspond to the CBP variables. An example of the former is a miscoded script that made it appear that House bills reported from a particular committee were being reported from a Senate committee. This sort of error that is hard to catch if you don’t know what you are looking for. But in this case, if you did a summary of House bill referrals (e.g.) you might notice something suspicious (etc.).

An example of the latter are the CBP variables PassS and PassH (passed Senate, passed House). There is not just one field that indicates bill passage. There are also fields for bills passed by unanimous consent or under suspension of the rules. So we had to collect information from multiple field to create complete PassH PassS variables. The good news is that Rohnin’s code should already do this.

**CBP variables created from multiple sources of information**. A good example is MREF (member of referral committee). To create this boolean we need to collect information about all of the committees a member serves on, and then all of the committees the bill was referred to. Then we have to write a conditional where if any of them match, then MREF =1. There are a number of other variables like this, such as CHREF (chair of referral committee). The good news is that all of this information is available from the github site now.

**Biographical information**. We originally collected this from the Database of Historical Congressional Statistics, but they stopped updating after 1996. We haven’t done a good job of maintaining this part of the dataset. The github site include scripts for a lot of biographical data, so this could be something to update. But just having the current database updated regularly will be greatly appreciated by a lot of researchers.

**Final comments!** I strongly recommend using congress.gov to double check the data before you release it (good task for undergraduates). The PassH, PassS example above illustrates why this is important. Someone noted that we were missing a lot of passed bills.

Unfortunately, it’s hard to check for every possible error. At the same time, be sure to continue to convey to researchers that we do our best but they should are responsible for checking the data.

There’s also the congressional bills.org website, which hasn’t been updated in years (except for the codebook and downloads). It may make sense to integrate into the CAP at some point.