GPO 113-116th Bill Status Bulk Data Repository Analysis

116th House Bills Bulk Data Location: <https://www.govinfo.gov/bulkdata/BILLSTATUS/116/hr>

Bill XML Page URL format:

* https://www.govinfo.gov/bulkdata/BILLSTATUS/116/hr + “/BILLSTATUS-116hrNNNN.xml”
  + where bills range in number from NNNN = 1 to 8578
  + HR 8578 was created 10 October 2020
  + Numbers in sequence may be skipped – it appears they may be 8 skips in the numbering of the bills and the files
* Example:  
  <https://www.govinfo.gov/bulkdata/BILLSTATUS/116/hr/BILLSTATUS-116hr43.xml>

Analysis outline:

* Create House Membership, “HouseMembers”, database table to store:
  + Index number
  + Last name of congressman
  + First name of congressman
  + State/Territory
  + District number
  + Party affiliation
* Parse House Clerk XML file to create database
  + Membership XML file can be found at: <https://clerk.house.gov/xml/lists/MemberData.xml>
  + Membership XML file was saved to (on 10/12/2020):  
    /Users/tim/Documents/PY4E/capstone/Data Analysis Project/116th Members/Office of the Clerk - MemberData - 116th.xml
  + Discard existing and create database from new in case there are any membership changes (e.g. filled vacancies) between crawls
  + Record extracted from XML file:
    - Last name of congressman = members/member/member-info/lastname
    - First name of congressman = members/member/member-info/firstname
    - Biography Guide ID = members/member/member-info/bioguideID
      * Unique identifier from the House Clerk
      * E.g. Andy Levin = L000592
    - State/Territory = members/member/statedistrict
    - District number = members/member/statedistrict
      * State/district is concatenated – e.g. “MI09” for Andy Levin
      * Will have to split string
    - Party affiliation = members/member/member-info/party
      * R = Republican, D = Democrat, L = Libertarian, I = third party
      * In program for data analysis, we will group all third parties into “I” for independent or 3rd party.
* Create, if not exists, and open (cur) House of Representative Bills, “HRBills”, database table(s) to create record:
  + Index number
  + Legislation type: HR (House Bill), S (Senate Bill), HJRES (House Joint Resolution), SJRES (Senate Joint Resolution), HCONRES (House Concurrent Resolution), SCONRES (Senate Concurrent Resolution), HRES (House Simple Resolution), SRES (Senate Simple Resolution)
  + Originating chamber: Senate, House, Joint
  + Legislative Number
  + Congressional Session
  + Legislation Title
  + Sponsor(s)
  + Co-sponsor(s)
* Create, if not exists, and open (cur2) “SponsorOriginators” database table
  + Many to many relationship
  + Multiple members can jointly introduce a bill as sponsors
  + Documents all sponsor to sponsor connections on introduced HR bills
  + If a bill only has a single sponsor, then no sponsor to sponsor connection is recorded
* Create, if not exists, and open (cur3) “SponsorSupporter” database table
  + Many to many relationship
  + Documents all sponsor to co-sponsor connections on introduced HR bills
  + *Resource: Review PY4E book example where teacher-student relationship was documented in a many-to-many table*
* Crawl through House bills for the 116th Congress and update database table
  + If there are five consecutive legislative numbers without an associated XML file, then assume that you have reached the end of the active legislation – in this case HR Bills.
  + Base\_URL = https://www.govinfo.gov/bulkdata/BILLSTATUS/116/hr
  + LegiType = /BILLSTATUS-
  + CongressNo = “116”
  + Chamber = “hr”
  + Article = “NNNN”
    - In crawl, ask user how many bills do they want to parse
      * Set up a countdown counter, “i”, based on the user response
      * Decrement counter each time through loop
    - In crawl, set up a counter, NNNN, that starts at:
      * 0 initially
      * Or max ID in “HRBill” database
  + URL = Base\_URL + LegiType + CongressNo + Chamber + Article + “.xml”
* For each HR Bill XML file:
  + Create record in “HRBills” database table
  + Write attribute in “HRBills” database table for:
    - Index number
    - Legislation type: bill, resolution = billStatus/bill/billType
    - Originating chamber: Senate, House, Joint = billStatus/bill/originChamber
    - Legislative Number = billStatus/bill/billNumber
    - Congressional Session = billStatus/bill/congress
    - Legislation Title = billStatus/bill/title
    - Sponsor(s) = billStatus/bill/sponsors/item/bioguideId
      * Note: case in XML tree – especially bioguideId – Last “d” is lower case
      * Note: there may be more than one sponsor
    - Co-sponsor(s) = billStatus/bill/cosponsors/item/bioguideId
      * Note: there may be more than one sponsor
  + Record any sponsor-to-sponsor relationships in “SponsorOriginator” table
  + Record any sponsor-to-co-sponsor relationships in “SponsorSupporter” table
* Create JavaScript Social Network visualization diagram based on:
  1. Sponsor to Co-sponsor connections on a bill
  2. Sponsor to Sponsor connections on a bill
  3. Co-sponsor to Co-sponsor connections on a bill

Missing HR Bills in Sequence:

* A committee will hold a "mark-up" session during which it will make revisions and additions. If substantial amendments are made, the committee can order the introduction of a "clean bill" which will include the proposed amendments. This new bill will have a new number and will be sent to the floor while the old bill is discarded. The chamber must approve, change or reject all committee amendments before conducting a final passage vote.
* In recent Congresses, the resolution specifying House internal rules of procedure includes reserving bill numbers for assignment by the Speaker. In the 112th Congress (2011-2012) the practice was extended to reserve additional bill numbers for assignment by the Minority Leader.
* *Try to read XML file, if FILE NOT FOUND, then create BLANK record for reserved or discarded bill numbers.*
* *As a result, id should equal legNo and there should be no skips in the database table, and every possible HR bill number s accounted for*
* *If this program were to be regularly used, I would create a program that would be periodically ran to search for “activated” bill numbers that were previously reserved or discarded.*

Program performance

* Accessed GPO Bulk Data website, and parsed 500 HR Bills XML files:
  + - 8:25AM – 2 min 45 sec
  + Size of 116-HR-net.sqlite
    - With 8725 HR bills parsed into HRBills table and 25 bills parsed into Sponsor table: 98.4 MB
    - With 8725 HR bills parsed into HRBills table and 8725 bills parsed into Sponsor table: 102.8 MB
  + Time to parse 1000 HRBills tale records into Sponsors table: 18.37s
  + Time to parse 5000 HRBills tale records into Sponsors table: 1:27.58 min:sec