The Devil You Know: Threat Construction in the United States

1

DATA SOURCE:
INTERNET ARCHIVE'S .GOV COLLECTION

E-SCIENCE DATA INCUBATOR PARTICIPANT: EMILY KALAH GADE

PROJECT LEADS:
DR. JOHN WILKERSON
DR. MAGDALENA BALAZINSKA

Project Background: Scope

- 2
- Political Science Side NSF Grant re: Poli-Infomatics
 - Explore the wealth of textual resources available today
 - ▼ Measuring ideas in text... Using computers to do it
- Computer Science Side Text in Myria

Project Background: Data Source



- Internet Archive .gov collection
 - o 90 terabytes of data, 1.1 billion captures
 - Hosted by Altiscale on a Hadoop Cluster
- Incubator project:
 - o Figure out what is "in there"
 - Make it useful for social science research
 - ▼ Focus: Language about threats terrorism, climate change, and banking crisis

Incubator Goals

4

- Access the cluster, run scripts, get output
- Reliability and validity
- Results and write up
- Make it reproducible and accessible to other social scientists

Initial Hurdles

- Emily's coding abilities
- General big data issues
- Nit picky issues: unicode, parsers, carriage returns
- Best format for examining data
- Reliability and validity
- Reproducibility for anyone else (without the incubator!)

What a capture looks like once parsed

 $gov, senate, src)/public/\underline{index.cfm}?\& \underline{cfid}=44361862\&\underline{fuseaction} = senators.television \&\underline{istextonly} = false\&senator_id=34$ http://src.senate.gov/public/index.cfm? FuseAction=Senators.Television&Senator_id=34&CFId=44361862&IsTextOnly=False&CFID=29266014&CFT0KEN=92197829 20070306000548 Senate Republican Conference :: Senators Senate Republican Conference :: Senators Senate Republican Conference Chairman Kyl | Republican Senate gov | RPC | Vice Chair Cornyn | C-SPAN Live | DoD Video Briefings Home | Graphics Version About the SRC | Committee Assignments | Television | Radio | Graphics | Pen & Pad | Contact Search: Education Energy Global War on Terrorism Health Care Immigration/Border Security Jobs and the Economy Judicial Nominations Tax Relief United States Senator Jim Inhofe, one of the leading conservative voices in the Senate, is a strong advocate of common sense Oklahoma values including less government, less regulation, lower taxes, fiscal responsibility and a strong national defense. He was first elected to the Senate in 1994 to complete the unexpired term of Senator David B... Latest Television clips Browse by: -- Month -- January February March April May June July August September October November December -- Year -- 2002 (6) 2003 2004 2005 (1) 2006 (10) 2007 (15) February 2007 14th Senator James Inhofe Addresses BRAC Shortfalls on Senate Floor before Continue Resolution Vote Senator(s): James Inhofe Issue(s): Budget, Defense File name: inhofe cr feb14.ram Play Television clip 14th Senator James Inhofe Remarks on BRAC Spending Shortfalls Senator(s): James Inhofe Issue(s): Budget, Defense File name: inhofe brac feb14 ram Play Television clip 12th Senator James Inhofe talks about businesses and climate change at Environmental and Public Works Committee Hearing Senator(s): James Inhofe Issue(s): Environment File name: inhofe feb12.ram Play Television clip 9th Senator James Inhofe Opening Statement from EPW Oversight Senator(s): James Inhofe Issue(s): Environment, Environment/Natural Resources File name: epw_feb9.ram Play Television clip 8th Senator James Inhofe Speaks on the Senate Floor about the BRAC Spending Shortfalls Senator(s): James Inhofe Issue(s): Defense File name: inhofe feb8.cam Play Television clip 1st Senator James Inhofe questions General George Casey at his SASC Nomination Hearing Senator(s): James Inhofe Issue(s): febl casey ram Play Television clip 1st Senator James Inhofe Discusses Global Warming on Larry King Live Senator(s): James Inhofe Issue(s): Energy, Environment File levision clip January 2007 31st Senator James Inhofe Discusses Global Warming with Miles O'Brien on CNN's American Morning Senator(s): James Inhofe Issue(s): an31 obrien ram Play Television clip 31st Senator James Inhofe and Glen Beck Discuss Global Warming Senator(s): James Inhofe Issue(s): Environment File name: Environment vision clip 30th Senator James Inhofe Questions Admi wices Committee Senator(s): James Inhofe 🛽 evision clip Real Audio/Video (ram) Assistance: Rea hat allows you to view and listen to s stalled on your computer to access Real formated files. information and/or to download Real Player. pov,senate,energy)/public/index.cfm?displaydate=07/25/05&fuseaction=hearings.bymonth http://energy.senate.gov/public/index.cfm?FuseAction=Hearings.ByMonth&DisplayDate=07/25/05 20070207110337 U.S. Senate Committee on Energy and Natural Resources | 200 U.S. Senate Committee on Energy and Natural Resources United States Senate Committee on Energy and Natural Resources Home | Search | Graphics Site Map | FAQs About the Committee » Overview » History » Jurisdiction » The Ranking Member Schedule Printed Hearings Legislation Major Issues » Bill Status » Issues Senate Schedule News Room wublican Press Releas Conferences Contact Information Related Links » Legislation Federal Agencies Live Webcast Hearings Note: All Press Rel August 2005 » Full List Jul 12 at 10:00 AM Full Committee Nom n our website. Ju « June 2005 To consider the nominations y of Energy for Congressional and Intergovernmental Affairs; David R. Hill, of Missouri, to 🛭 ming, to be Assistant y of Energy for Environmental Management. Jul 12 at 03:00 PM Subcommittee on Water and Power joint Federal-State Floodplain and Erosion Mitigation Commission for the State of Alaska; S. 247, to authorize the Secretary of the Interior to ass construction of the Tumalo Irrigation District Water Conservation Project in Deschutes County, Oregon; S. 648, to amend the Reclamation States Emergence the authority for drought assistance; S. 819, to authorize the Secretary of the Interior to reallocate costs of the Pactola Dam and Reservoir, South Dakota, to reflect increased demands for municipal, industrial, and fish and wildlife purposes; S. 891, to extend the water service contract for the Ainsworth Unit, Sandhills Division, Pick-Sloan Missouri Basin Program, Nebraska; and S. 1338, to require the Secretary of the Interior, acting through the Bureau of Reclamation and the United States Geological Survey, to conduct a study on groundwater resources in the state of Alaska, and for other purposes. Jul 14 at 10:00 AM Full Committee Nomination Hearing To consider the nominations of R. Thomas Weimer to be Assistant Secretary of the Interior; and Mark Limbaugh to be an Assistant Secretary of the Interior. Jul 14 at 02:30 PM National Parks Subcommittee To review the National Park Service's business strategy for operation and management of the national park system, including development and implementation of business plans, use of business consultants, and incorporating business practices into day-to-day operations. Jul 14 at 03:00 PM Conference - HR 6, Energy Policy Act of 2005 Conference - HR 6, Energy Policy Act of 2005 Conference - HR 6, Energy Policy Act of 2005 Conference - HR 6, Energy Policy Act of 2005, webcast at: http://energycommerce.house.gov/108/Mark ___ul 19 at 02:30 PM Full Committee Hearing Oversight hearing to receive testimony regarding the effects of the U.S. nuclear testing program on the Marshall Isl on Public Lands and Forests To receive testimony on S. 703, to provide for the conveyance of certain Bureau of Land Management land in the Sta and for other purposes; S. 997, to direct the Secretary of Agriculture to convey certain land in the Beaverhead-Deerlodge Forest, Montana, to Je y; S. 1131, to authorize the exchange of certain Federal land within the State of Idaho, and for other purposes; S. 1170, to establish the Fort S n Area; S. 1238, to amend the Public Lands Corps Act of 1993 to provide for the conduct of projects that protect forests, and other purposes; and er with respect to certain lands erroneously included in the Cibola National Wildlife Refuge. California. Jul 21 at 10:00 AM Full Committee Hearing ics" To receive testimony regarding the current state of climate change scientific research and the economics of strategies to manage climate change. Issues to be discussed include: the relationship between energy consumption and climate change, new developments in climate change research and the potential effects on the U.S. economy of climate change and strategies to control greenhouse gas emissions. Note: Due to time constraints, Panel 2 has been postponed and escheduled for September 20th at 10:00 a.m. Jul 21 at 10:00 AM Full Committee Business Meeting To consider pending nominations. (The Full Committee hearing regarding Climate Change will commence immediately following the Business Meeting.) Jul 21 at 01:00 PM Conference – HR 6, Energy Policy Act of 2005 Conference – HR 6, Energy Policy Act of 2005 (for web access: http:// energycommerce.house.gov/108/Markups/07212005markup1601.htm) Jul 24 at 02:00 PM Conference - HR 6, Energy Policy Act of 2005 - http://energycommerce.house.gov This is to advise you that the next meeting of the Energy Bill Conference will be on Sunday, July 24 at 2:00 p.m. in Room 2123 Rayburn. Jul 25 at 02:00 PM Conference – HR 6, Energy Policy Act of 2005 – http:// energycommerce house gov Meetings of the House and Senate Conferees on H.R. 6, The Energy Policy Act of 2005 Jul 27 at 03:00 PM Energy Subcommittee Hearing – Hydrogen & Fuel Cell Technology to receive testimony on recent progress in hydrogen and fuel cell research sponsored by the Department of Energy and by private industry. Testimony will also address the remaining challenges to the development of these technologies. Jul 28 at 10:00 AM Subcommittee on National Parks Hearing - Miscellaneous Parks Bills To receive testimony on the following bills: S. 584 and H.R. 32, bills to require the Secretary of the Interior to allow the continued occupancy and use of certain land and improvements within Rocky Mountain National Park; S. 652, a bill to provide inancial assistance for the rehabilitation of the Benjamin Franklin National Memorial in Philadelphia, Pennsylvania, and the development of an exhibit to commemorate the 300th anniversary of the birth of Benjamin Franklin; S. 958, a bill to amend the National Trails System Act to designate the Star-Spangled Banner Trail in the States of Maryland and Virginia and the District of columbia as a National Historic Trail; S. 1154, a bill to extend the Acadia National Park Advisory Commission, to provide improved visitor services at the park, and for other purposes; S. 166, a bill to extend the authorization of the Kalaupapa National Historical Park Advisory Commission; and S. 1346, a bill to direct the Secretary of the Interior to conduct a study of maritime sites in the State of Michigan. Committee on Energy and Natural Resources 364 Dirksen Office Bldg 🛭 Washington, DC 20510 202.224.4917 Home | Search | Graphics On | Site Map

What word count data looks like



```
desertification 1
                dot\.gov
2009
        5
2009
                hhs\.gov
        5
                                  (fresh\swater) 21
2009
        5
                state\.qov
                                 desertification 1409
        5
                                 (climate\schange)
2009
                state\.gov
                                                           28204
        5
                                 pollution
2009
                energy\.gov
                                                  5812
        5
2009
                energy\.gov
                                 anthropocene
                                                  1
        5
                                         655546606
2009
                \.house\.gov
                                 total
2009
        5
                defense\.gov
                                 pollution
        6
                ed\.qov pollution
2009
                                          732
                va\.qov (global\swarming)
2009
                                                  3
2009
        6
                dod\.qov
                                 (food\ssecurity)
                                                           3
                                                           274
2009
        6
                dod\.gov
                                  (natural\sdisaster)
2009
        6
                dol\.gov
                                  (forest\sconservation)
                                                           9
2009
        6
                                  (food\ssecurity)
                                                           25
                dot\.qov
                                 (greenhouse\sgas)
2009
        6
                dot\.gov
                                                           941
2009
        6
                hhs\.gov
                                 pollution
                                                  1151
2009
        6
                state\.qov
                                 (food\ssecurity)
                                                           4959
2009
        6
                state\.qov
                                 (natural\sdisaster)
                                                           9621
        6
                \.senate\.gov
                                 anthropogenic
2009
                                 (ocean\sacidification)
2009
        6
                \.senate\.gov
                                                           48
2009
                whitehouse\.gov (global\swarming)
                                                           281
                                                  138771
        7
                         (greenhouse\sgas)
2009
                 other
2009
        7
                va\.gov (climate\schange)
                                                  16
2009
        7
                dod\.gov
                                 (fresh\swater)
                                                  55
        7
                state\.qov
                                 (fresh\swater) 838
2009
2009
        7
                energy\.gov
                                 (forest\sconservation)
                                                           2
2009
        7
                \.house\.gov
                                 (security\sof\sfood)
                                                           147
2009
        7
                \.senate\.qov
                                 total
                                          545452207
                whitehouse\.gov desertification 1
2009
        7
2009
        7
                whitehouse\.gov (climate\schange)
                                                           2157
                va\.qov total
                                 100787717
2009
        8
                va\.gov (natural\sdisaster)
2009
        8
                                                  492
```

Data Quality

- 8
- Complete Crawls Taken from Nov to Jan around elections from 2004 to present
 - o This becomes Gold Standard

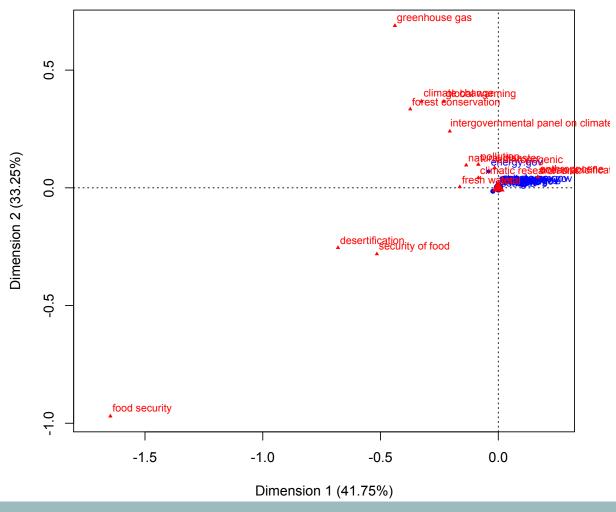
- Early data crawled irregularly
 - Early data sensitive to terms

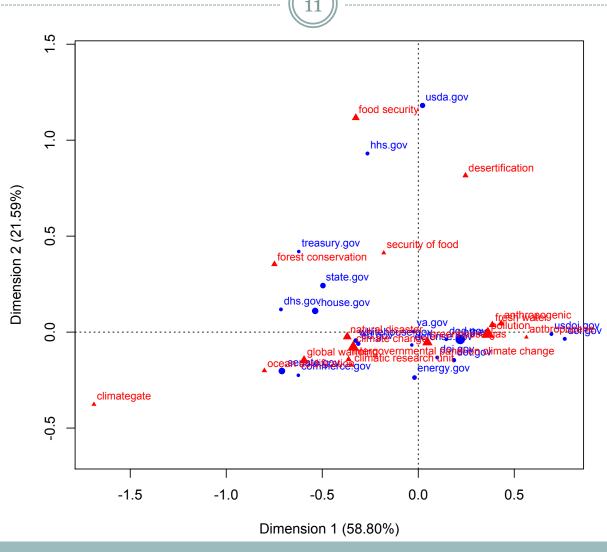
Preliminary Results

Correspondence Analysis

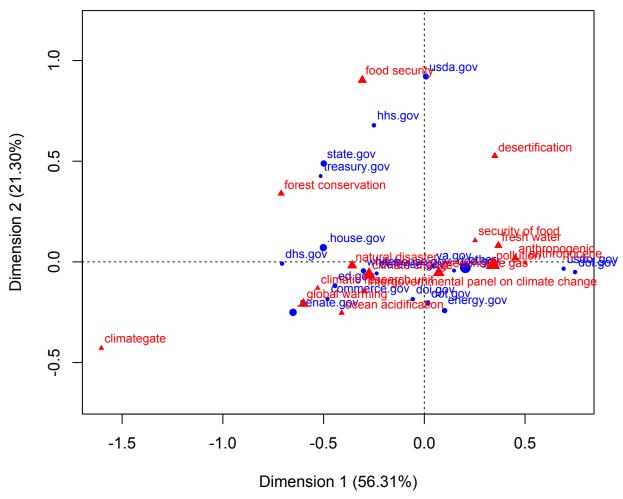
- Angles between topics and agency indicate more emphasis of that topic by that agency (smaller angels = more emphasis)
- Interagency distances and inter topic distances are interpretable (but not topic-agency distances)

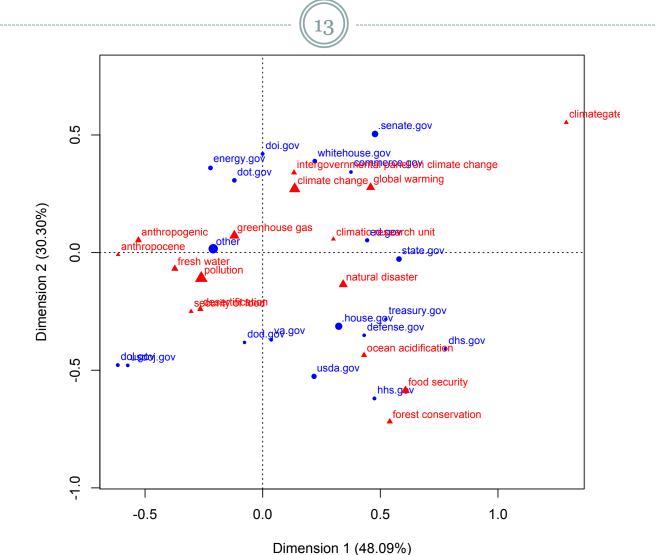












Creating Institutional Memory



- Taking data to PoliInfomatics conference
- Publishing code on Github, publishing data on PoliInfomatics
- Contact with Internet Archive and two other groups known to be exploring this data
 - One at Harvard, one in Germany
- Article re: .gov collection as a data source

Proposed Publications



- Methodological Paper re: .gov data as a data source for social science research
 - Including preliminary results and scripts
- The Devil You Know: Threat Construction in the United States
 - Tracking threat construction in the US; Agenda analysis

Thank You!