Victoria Mitchell, UO Libraries
Joshua Baker, Cary Fontana and Shawna Meechan, Political Science Department
University of Oregon

Library & Researcher Collaborative Data Publishing:

THE SOUTHERN VOTING PROJECT

Global Environment

- Barriers to social science researchers sharing data
  - Researcher reluctance
- “Small” or “long tail” data
  - Need support from institution
  - Most social science data = small

Local Environment - UO

- No campus research data policy
- Library leading on research data management and preservation
  - Collaborating with campus IT, Research Services
  - Digital Scholarship Center
    - Open Access Publishing
    - Digital Collections
    - Institutional Repository
    - Interactive Media Development
  - Data Services
    - Science Data Services Librarian
    - Recently created (partial) position of Social Science Data Librarian

Photo credit: University of Oregon Libraries

Matching up needs…

Political Science professor

- Junior faculty member doing quantitative research
- Looking for funding help to finish collecting data for analysis
- Working on 3 different data-heavy projects

- Library
  - Wants to get more data sets into the IR
  - Wants to test a data publishing project; not just IR
  - Open to a pilot project using library “initiative” funds

Photo credit: University of Oregon Libraries

Planning

Made proposal to Library Council for funding project

- Aligns with library strategic directions
- Interested parties communicate and come up with a proposal
  - Propose budget of $6000 – for graduate Research Assistants (RAs)
  - Timeline of June 2013-January 2014
  - Keep technology to what is free or already available to keep cost down

Pre-Project Launch

- Two potential RAs – 2 political science grad students who had taken a pilot research data management course in the spring

Decisions

- Data to be collected
  - Pilot project -- limit to data for one of the research projects
  - Chose: African-American enfranchisement and policy outcomes 1938-2012
- Chose platform/software for making data and maps available
  - Create website using Wordpress – Campus already has support through EduBlogs and central Information Services
  - Create custom maps using GeoCommons – Online open access data repository
- Create codebooks for metadata

Logistics

- Time frame
  - Start date – June 17 (first Monday after spring finals week)
  - End date – September
- Hire graduate student research assistants (RAs)
  - 3 RAs – Cary Fontana, Josh Baker, Shawna Meechan
  - One Lead RA (Fontana)
  - 50 hours per week – 30 for lead RA, 10 each for other two RAs
- Consulting library team partners
  - Victoria Mitchell, Social Sciences Data Librarian
  - Karen Estlund, Head, DSC
  - John Russell, Scholarly Communications Librarian
  - Azle Malinao-Alvarez, Instructional Technology Consultant

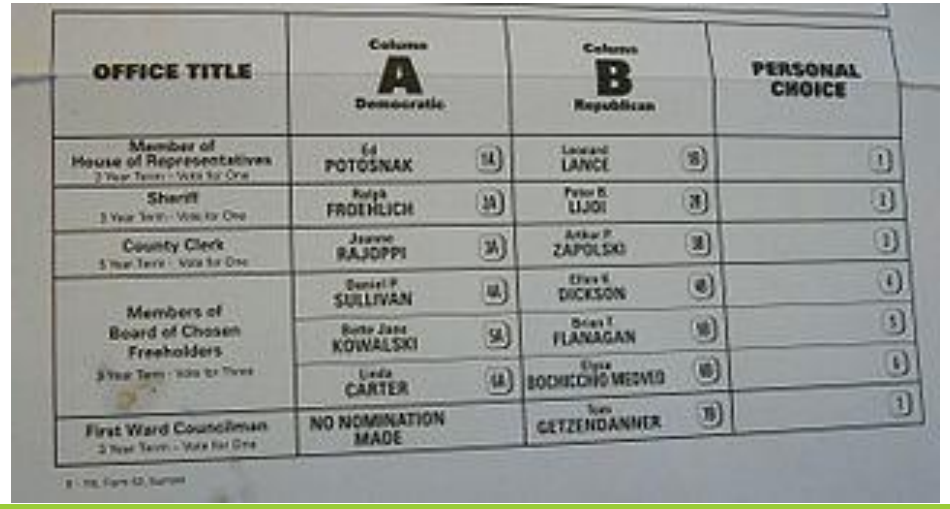
Data collected and curated

- Political and Demographic Data
- Wanted variables on state, Congressional District, and county levels
  - 3 separate datasets by geographic level
- Eventual total of 335 variables included:
  - 136 State
  - 100 County
  - 99 Congressional District

Political & Policy variables

- Voter registration rates by race
- Other registration & voting variables
- State and county policy outcomes

- Election results by state, CD & county for:
  - President, Senate, House, Governor
- Congressional district outcomes



OFFICE TITLE	Column A Democratic	Column B Republican	PERSONAL CHOICE
Member of House of Representatives 2 Year Term - Vote for One	Ed POTOSNAK 14	Leonard LANCE 15	1
Sheriff 3 Year Term - Vote for One	Kolya FROEHLICH 3A	Peter B. LIJOI 3B	1
County Clerk 5 Year Term - Vote for One	Jeanne RAJOPPI 3A	Arthur P. ZAPOLSKI 3B	1
Members of Board of Chosen Freeholders 3 Year Term - Vote for Three	Daniel P. SULLIVAN 4A	Elias K. DICKSON 4B	1
	Bette Jane KOWALSKI 5A	Brian T. FLANAGAN 5B	1
	Linda CARTER 6A	Elyse BODNICZKO WEGNER 6B	1
First Ward Councilman 3 Year Term - Vote for One	NO NOMINATION MADE	Sam GETZENDANNER 7B	1

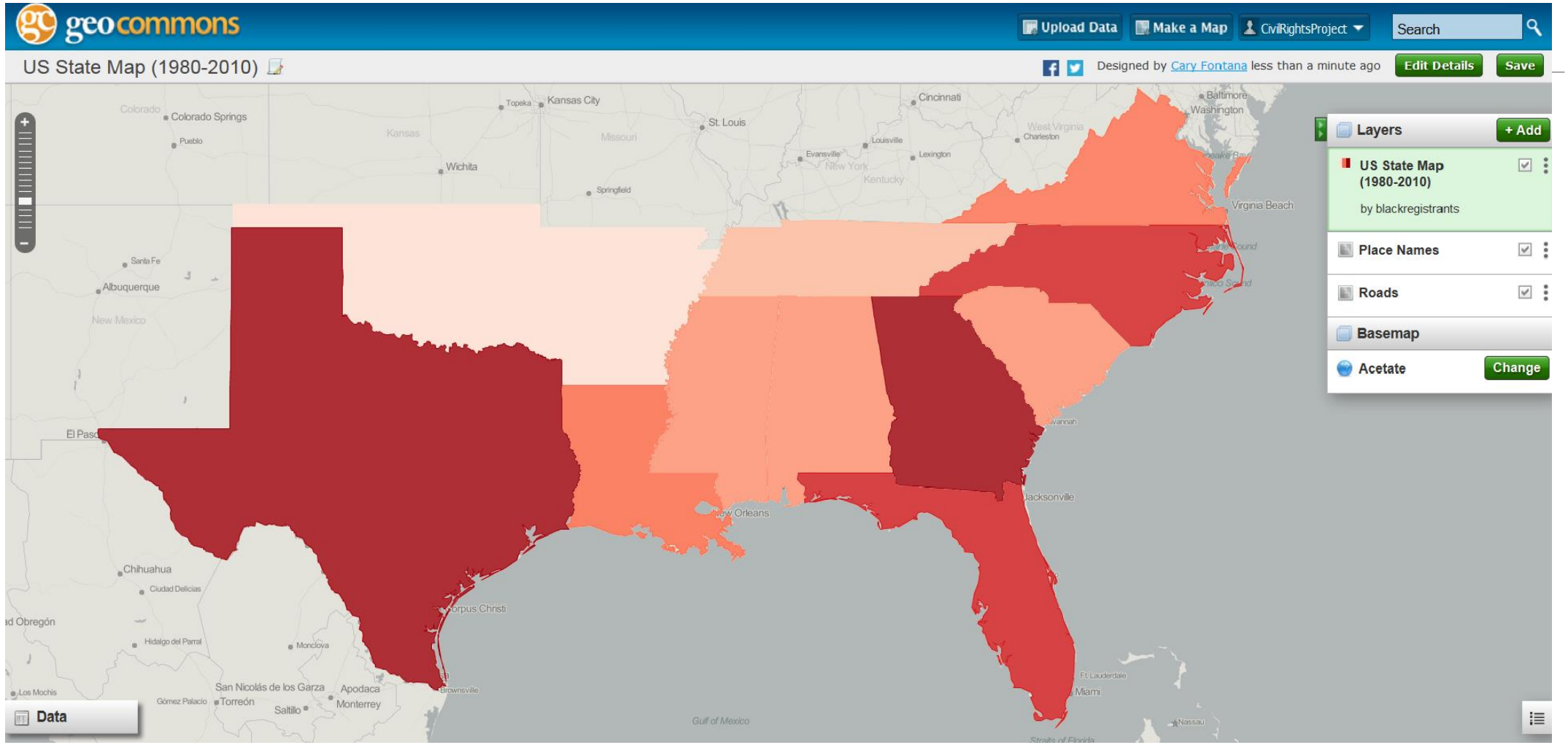
Demographic (Census) variables

- Total population
- Percent population African-American
- Unemployment rate
- Percent of population aged 65+
- Percent of population aged 5-17
- Percent of families below poverty line
- Child poverty rate
- Percent adults with high school degree
- Per capita income
- …plus others

State-level data

- Over 600 total variables available, 136 used
- Most interested in the budgetary variables describing the sources of state funds and how they were spent
- Data sources: Multiple U.S. Census Bureau publications including “Census of Governments,” “Current Population Survey” (CPS), “State Government Finances”; U.S. Statistical Abstract.

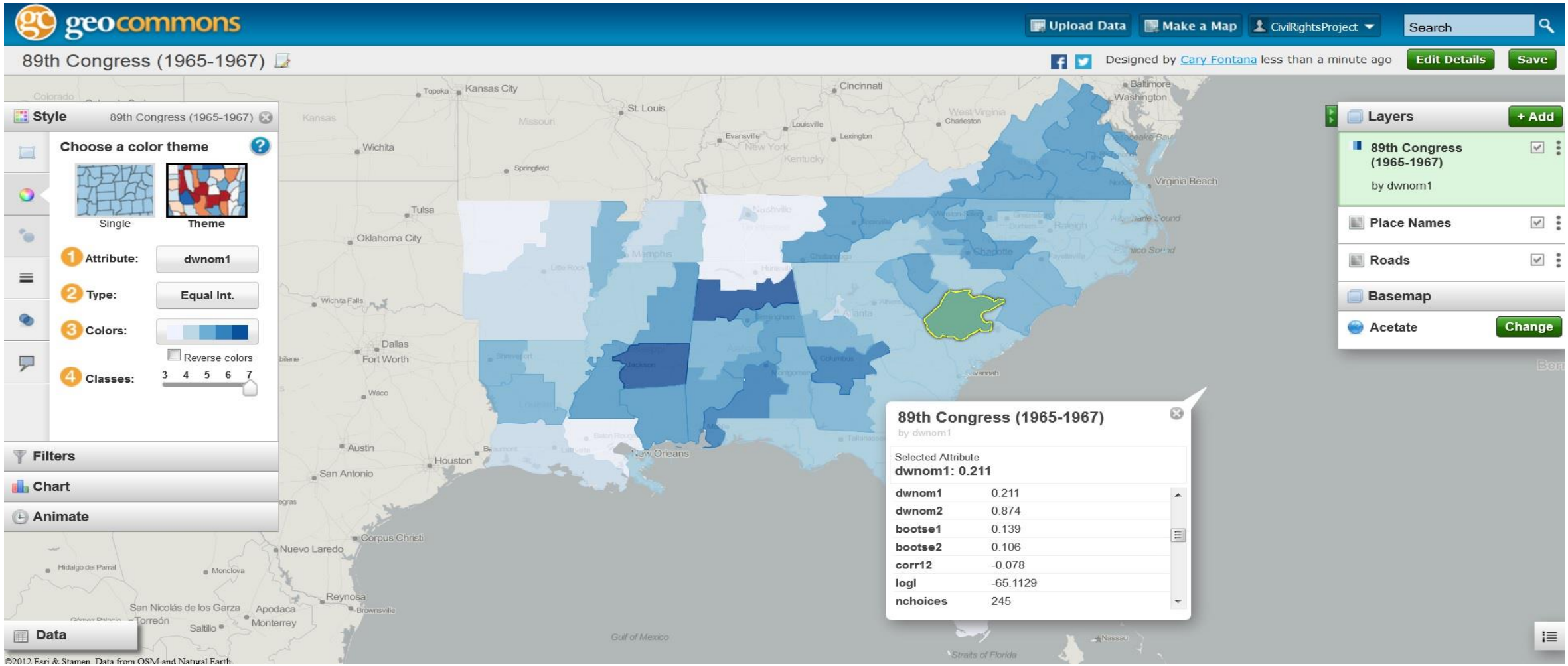
State-level Map



Congressional District Variables

- Collected and organized 99 Congressional District Level Variables:
  - Data on White and African-American Registrants (circa 1958-2012) from the 11 former states of the confederacy was compiled by our research team.
    - Registration data was mined from State Archives, Civil Rights Commission Reports, and Secretary of State's offices .
  - Other data was compiled from various data sources.
- Temporal Coverage:
  - Much of our data covers 1946-2008.
  - Some data extends back to the First Congress (1789) through 2008.

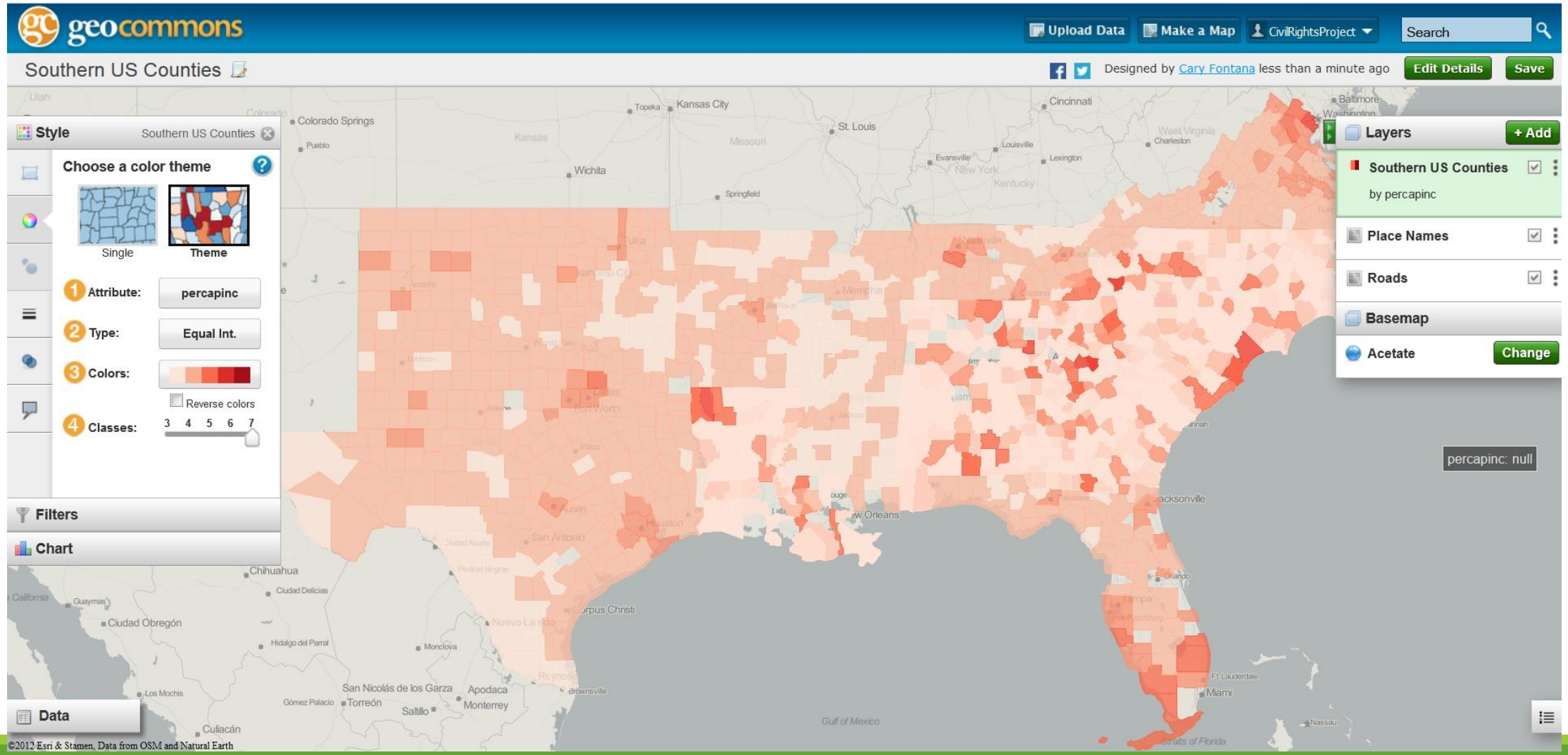
Congressional District Map



County-level Variables

- Collected and organized demographic data at the county level for each state.
  - Data was collected concerning race, age, and income of multiple measures for individuals making up the population for each county nationwide.
  - Data was also collected concerning education levels, employment levels, and poverty information for families and individuals.
  - Data was compiled using census data made available through various sources
- Temporal Coverage:
  - All data covers 1970-2010
  - Large portion of county data stretches back to 1940.

County-level Map



The Southern Voting Project Website

Home | About | Data | Maps | Publications/Working Papers | Acknowledgements

Home › About

About

Welcome to the Southern Voting Project. The project, led by Dr. William Terry of the University of Oregon political science department in conjunction with the University of Oregon Library's Digital Scholarship Center, presents the most comprehensive dataset to date pertaining to the politics of the US South. On this site you will find links to the data and a set of interactive mapping tools hosted by GeoCommons.

Description of data

The data cover three levels of analysis: (1) state, (2) US Congressional district, and (3) county level. Data for each level is available on this site in *.csv* and *.dta* format with a codebook and a *pdf* of the original source material.

There are three broad categories of variables in our datasets: (a) political outcomes, (b) policy outcomes, and (c) demographic variables. Category (a) includes variables dealing directly with politics, e.g., election results and the number of voters in the focal jurisdiction. Category (b) contains information about policy outcomes, such as county-level education expenditures, the state corporate income tax rate, and the roll call voting record of the district's member of Congress. Category (c) contains standard demographic information about the focal jurisdiction, e.g., its racial composition, per capita income, percent of the population elderly, incarceration rates, etc.

Data Preparation Activities

- Lead RA trained on GeoCommons, WordPress, and on how to deposit into Scholars' Bank (DSpace based-institutional repository)
- Data files cleaned up in Stata; preserved and deposited in both Stata and comma-delimited formats.
- Data documented in 3 codebooks (state, Congressional district, county) in Excel format

Outputs

- 3 data sets, with codebooks, deposited into Scholars' Bank
- "Southern Voting Project" web site created
- Maps created via GeoCommons and embedded into web site

Issues & Lessons Learned - Faculty & RA's

- RA's
  - Talk more to each other
  - Stata learning curve
  - Choosing best data sources at outset (Consult experts)
  - Converting pdfs to numerical data formats – are there technological solutions?

Issues & Lessons Learned - Libraries

- We can scale:
  - Institutional Repository Capacity
- To a lesser extent, we can scale
  - Expertise in Library to Help with metadata (5 hours)
  - Basic web / WordPress tutoring (5-10 hours)
  - Help with existing tools like GeoCommons (5-10 hours)
- Costs: $10K + staff commitments
- Free interactive tools are problematic to develop with embargoed data
- Data licensing and copyright issues

References

1. Akers, Katherine G. 2013. “Looking Out for the Little Guy: Small Data Curation.” ASIS&T Bulletin February/March 2013.” *Bulletin of the American Society for Information Science & Technology*. http://www.asis.org/Bulletin/Feb-13/FebMar13_RDAP_Akers.html.
2. Borgman, Christine L. 2012. “The Conundrum of Sharing Research Data.” *Journal of the American Society for Information Science and Technology* 63 (6): 1059–1078. doi:10.1002/asi.
3. Cragin, Melissa H, Carole L Palmer, Jacob R Carlson, and Michael Witt. 2010. “Data Sharing, Small Science and Institutional Repositories.” *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 368 (1926) (September 13): 4023–38. doi:10.1098/rsta.2010.0165. http://rsta.royalsocietypublishing.org/content/368/1926/4023.short.
4. Heidorn, P Bryan. 2008. “Shedding Light on the Dark Data in the Long Tail of Science.” *Library Trends* 57 (2): 280–299. doi:10.1353/lib.0.0036. http://muse.jhu.edu/content/crossref/journals/library_trends/v057/57.2.heidorn.html.
5. Wicherts, Jelte M, Marjan Bakker, and Dylan Molenaar. 2011. “Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results.” Edited by Rochelle E. Tractenberg. *PloS One* 6 (11) (January): e26828. doi:10.1371/journal.pone.0026828. http://dx.plos.org/10.1371/journal.pone.0026828.

? Questions ?

Victoria Mitchell, vmitch@uoregon.edu