The Pennsylvania Policy Database Project: A Model for Comparative Analysis

Joseph P. McLaughlin, Paul Wolfgang, J. Wesley Leckrone, Justin Gollob, Jason Bossie, Jay Jennings and Michelle J. Atherton

State Politics & Policy Quarterly 2010 10: 320
DOI: 10.1177/153244001001000306

THE PRACTICAL RESEARCHER

The Pennsylvania Policy Database Project: A Model for Comparative Analysis

Joseph P. McLaughlin, *Temple University*
Paul Wolfgang, *Temple University*
J. Wesley Leckrone, *Widener University*
Justin Gollob, *Mesa State College*
Jason Bossie, *Carnegie Mellon University*
Jay Jennings, *Temple University*
Michelle J. Atherton, *Temple University*

ABSTRACT

Temple University led a six-university effort that built a comprehensive public policy database for Pennsylvania, modeled on the national Policy Agendas Project created by Frank Baumgartner and Bryan Jones (1993). The Pennsylvania database (www.temple.edu/papolicy) enables users to integrate data from all three branches of government and the news media organized into 20 major and 249 minor policy topics since 1979. This article discusses the value of these data, their potential uses in state policy research, and the lessons learned over the four years invested in building the database. Our hope is that interested readers might undertake similar projects in their states to create a standardized national network of state policy databases.

RESEARCHERS ARE WELL aware of the challenges in conducting longitudinal analyses of public policy in U.S. state legislatures. Two of the many challenges deserve special attention. The first is data access or information retrieval. The lack of a consistent and comprehensive historical record makes tracing policy change across venues particularly difficult. The second challenge is data analysis or pattern recognition. Even when presented with a comprehensive historical legislative record, longitudinal and, for that matter, cross-sectional analyses are difficult to carry out given the lack of a standardized coding framework. A

State Politics and Policy Quarterly, Vol. 10, No. 3 (Fall 2010): pp. 320–336

database, which was constructed by scholars and students at Temple and five other universities on behalf of the Pennsylvania General Assembly, addresses these problems. This article introduces the dataset, discusses its utility, and shares the lessons learned during its construction in the hopes of engendering confidence in researchers hesitant to construct similar databases. The project also suggests that state governments themselves might be a new source of funding for policy-related research, particularly the reorganization of state archives to improve public accessibility and simultaneously meet social science standards for categorizing data.

ORIGINS AND PURPOSES OF THE PROJECT

In 2005, the Pennsylvania General Assembly agreed to fund the construction of a comprehensive policy database in response to a proposal from Temple University. In securing funding, Temple argued that the database would broadly benefit state legislators, other state policymakers, the press, the public, students, teachers, and academic researchers. Decision-makers would benefit by having a comprehensive historical record of policy problems and solutions. The press and public would benefit by having easier access to policy records, thereby satisfying a growing appetite for increased transparency. Teachers, students, and researchers would benefit by having a sound social-scientific tool of investigation to research state government and public policy. Finally, the legislature, whose caucus-based policy activities and communications with the public often—and rightly so—reflect conflict, would benefit from a comprehensible "no-spin" history that tells the larger institutional story.

To achieve these multiple benefits, Temple modeled its database after the U.S. Policy Agendas Project constructed by Frank Baumgartner and Bryan Jones (www.policyagendas.org). The design calls for collecting a variety of datasets—newspaper articles, legislation, committee hearings—and coding them into a single standardized policy topic schematic. This design creates a website that not only centralizes dispersed data effectively, but it also allows for the measurement of relative attention to policies across long periods of time.

The Policy Agendas Project was a suitable model for several reasons. First, the national model effectively prescribes a method of how to reconstruct legislative history that is both comprehensive and amenable to longitudinal analysis. This method has effectively overcome the main problems facing policy observers, as described above. Second, because state governments address many of the same issues the federal government faces, the Policy Agendas Project model was replicable at the state level. Finally, by adapting the Policy Agendas codebooks to Pennsylvania government, and by using the same decision rules for coding records, the Pennsylvania Policy Database can

be plugged into existing international, (Comparative Policy Agendas Project[1]) national (U.S. Policy Agendas Project and Congressional Bills Database[2]), and future comparable databases. If the Pennsylvania Policy Database Project is replicated in additional states, the result would be a network of state databases useful for studies in comparative state analysis, as well as federalism.[3]

USES OF THE DATABASE

The key feature of the Pennsylvania database is the ability to access standardized data relating to state policy since 1979 through a single, free, online source. The database is comprised of more than 157,000 records (not including fiscal data) in a number of important datasets, including state legislative hearings, bills, acts and resolutions, governors' budget addresses and executive orders, Supreme Court decisions, news stories, and state public opinion polls (See Table 1). All records are coded into one of 20 major and almost 250 minor topics based on their policy purpose.

Given the standardized coding, data are directly comparable across policy venues (e.g., legislative hearings, governors' budget messages, and public opinion polls), time, and topics (e.g., higher education).

The Pennsylvania Policy Database Project has a number of advantages over traditional government archives and similar resources. The first advantage is that of easier data access and retrieval. Users have access to decades of digitized records now dispersed in various institutional repositories in a single, easy-to-use online database. Because all datasets are downloadable, users have access to raw data, including the full text (as opposed to abstracts) and history of more than 75,000 bills with tens of thousands of amendments. This is particularly important to researchers who are interested in specific individual records. Finally, data retrieval has been improved and simplified by the standardized coding system. In practice, this means that users can search policy topics with confidence that they are not missing key records.

A second advantage of the project is the ability to analyze attention given to policy issues across various institutions from 1979 to the present. Unlike most government archives, which often rely heavily on key word search tools and typically code records into multiple categories, our database codes each record only once, into an exclusive and exhaustive set of policy topics, thereby defining for each dataset the total measurable policy space. Thus, our database provides users not only a more efficient form of information retrieval (as explained above) but also a function other databases cannot easily provide: pattern recognition. Because each record has been coded only once, users can graph the relative attention of legislators (bills, resolutions, legislative hearings,

Table 1. Datasets for the Policy Agendas Project and the Pennsylvania Policy Database Project

Dataset Description	Policy Agendas Project, 1947–2004	Pennsylvania Policy Database Project, 1979–2006 (157,528 Records)
Legislative hearings	Congressional hearings	State legislative hearings *and legislative studies* (6,335 records)
Legislation	Public laws	Acts, *bills, and resolutions* (75,653 records)
Executive orders	Executive orders (President)	Executive orders (Governor) (212 records)
Executive messages	State of the Union addresses	Governors' budget messages (28 messages, 7,655 sentences)
Supreme Court decisions	U.S. Supreme Court decisions	Pennsylvania Supreme Court decisions (4,547 records)
Budget	Budget (authorizations)	Expenditures, *fiscal condition, bond ratings**
Media	*New York Times* index	State capital news digests (57,883 records)
Most important policy issues	*Congressional Quarterly*	*Governing* magazine (5,215 records)
Most important problem (opinion)	Gallup polls	State public opinion polls** (28 polls)

Note: Italics indicate Pennsylvania records with no Policy Agendas counterpart.

* The Pennsylvania database will include as a measure of fiscal condition general fund resources, expenditures, balances, and budget stabilization fund balances (rainy-day funds) for Pennsylvania and all 50 states as compiled for the annual *Fiscal Survey of the States* published by the National Governors Association and National Association of State Budget Officers.

** The Keystone Poll, currently known as the Franklin & Marshall College Poll, conducted by Floyd Institute for Opinion Research at Franklin and Marshall College

legislatively authorized studies), governors (executive orders, governors' budget messages), the Pennsylvania Supreme Court (Supreme Court decisions), media (state capital news digests, *Governing* magazine), and citizens (opinion polls) across topics and years. As noted above, because of its integrated datasets and standardized architecture, we believe the database could be used to address questions across a variety of fields and across multiple governments.

A RESEARCH EXAMPLE: ANALYZING HEALTH CARE AND EDUCATION

Even as a single-state database, the project will facilitate studies useful to both state policymakers and the discipline more generally. To take a simple example, suppose that a researcher wanted to explore the hypothesis that rising health care costs were consuming attention and funds at the expense of

state investments in education, despite gubernatorial, legislative, and public preferences to the contrary. Although a number of 50-state studies have found a negative association between Medicaid and education funding (e.g., Kane, Orszag, and Gunter 2003; Boyd 2005; Tandberg 2008), a case study could amplify the nature of the relationship and provide contextual understandings not easily achieved without the kind of in-depth research more feasible in a single-state or a carefully constructed set of case studies (Nicholson-Crotty and Meier 2002).[4] This project might involve quantitative and qualitative methods, including examining spending patterns and changes in state fiscal conditions over time, measuring the relative weight given to health and education across a variety of venues, and surveying legislators, executive branch officials, lobbyists, and advocacy groups. Before undertaking the project, a researcher might well want to conduct a preliminary scan to test its plausibility and to identify key variables, i.e., the appropriate time span to cover and, perhaps, which years appear to have been turning points (if any) and so would merit deeper investigation.

Using existing resources, the researcher would need to conduct searches on multiple websites. Looking at the General Assembly's website, which is the principal source for legislative variables, such as laws, bill introductions, and public hearings, the researcher would face a number of problems. First, the researcher would only be able to search one session at a time, which complicates longitudinal analysis. Second, research would require guessing where to find relevant data under alphabetically organized policy topics. Does a researcher look under "E" for "education" or "S" for "schools" or both, where many of the same bills—and some different bills—can be found? Finally, legislation is often listed not only in multiple topics but as many as 20 times under the same topic, making it virtually impossible for users to count the total number of bills in a specific policy area over a session or multiple sessions.

Then the researcher might want to visit the governor's website to identify relative attention to health care versus education in annual budget messages and executive orders. This would require downloading and reading all of these records and coding them by policy topics. (Of course, a thorough analysis would require coding not only these two topics, but also all of the documents to see whether attention to other issues appears to be inversely related to health care attention.) The researcher might also want to scan trends in public opinion polls, court decisions, and the news media, all requiring separate searches and coding efforts. News stories and editorials might be useful, not only as attention records (How much weight did politicians see the news media giving these two topics over time?) but also as event records (What were key actors saying about these issues?). The news media search could prove to be highly

dependent on which outlet was selected, as Pennsylvania has no dominant newspaper of record.

The Pennsylvania Policy Database Project's website vastly simplifies these tasks. We centralize multiple datasets into a single database, standardize coding procedures for all datasets across multiple years, and present results with user-friendly graphs and tables to aid analysis. Because all records are filed once in exclusive and exhaustive categories, users can display raw counts and percentage measures of activity across multiple policy topics with no double-counting. A researcher using our database could quickly scan education versus health spending from 1979 to 2004, as well as the relative strength of the general fund, a reflection of the impact of the business cycle on the state (See Figure 1). Similarly, he or she could scan the relative attention governors' budget messages gave to health versus education for the same period (See Figure 2).

Zeroing in on a single session or year, the researcher could determine very quickly that, for example, in 1999, education accounted for 10 percent of legislation, 18 percent of the governor's budget messages, 11 percent of news digest stories, and 1 percent of Supreme Court decisions. In addition, it was viewed as the most important problem facing Pennsylvania by 16 percent of the public (See Figure 3). Only the Supreme Court paid more attention to health issues.

By 2005, when the state general fund balance was lower and health care spending much higher than it had been five years earlier (Figure 1), health was dominating education in all of these policy venues, except public opinion,

Figure 1. Health vs. Education Spending as a Percent of Total Spending and Balance or Deficit as a Percent of the General Fund by Legislative Session, 1979–2004

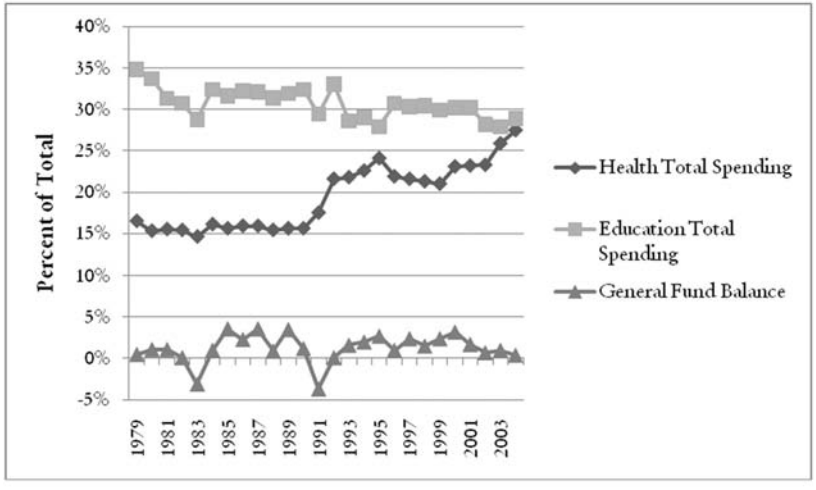


Figure 2. Governors' Budget Messages Devoted to Health vs. Education, 1979–2005

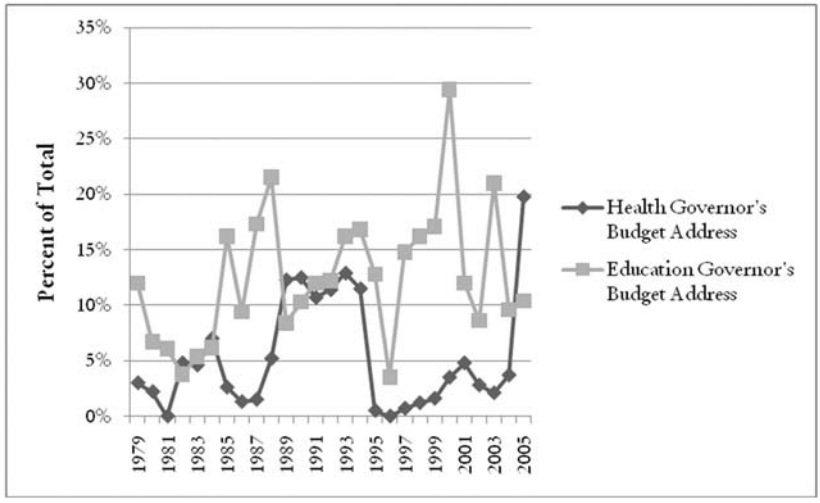
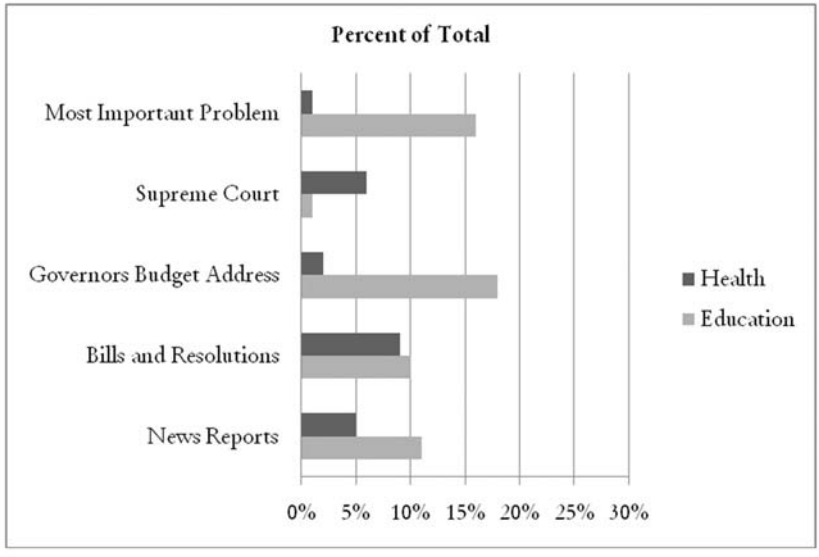


Figure 3. Attention to Health and Education Data in Selected Datasets, 1999



where education had a slight lead as a more important problem (See Figure 4). The researcher could then download selected datasets, reading very quickly what the governor had to say about these two topics in his budget message in any or all years since 1979 (See Table 2). Once the database scan is complete, the researcher will be able to proceed to a much stronger and better focused

Figure 4. Attention to Health and Education Data in Selected Datasets, 2005

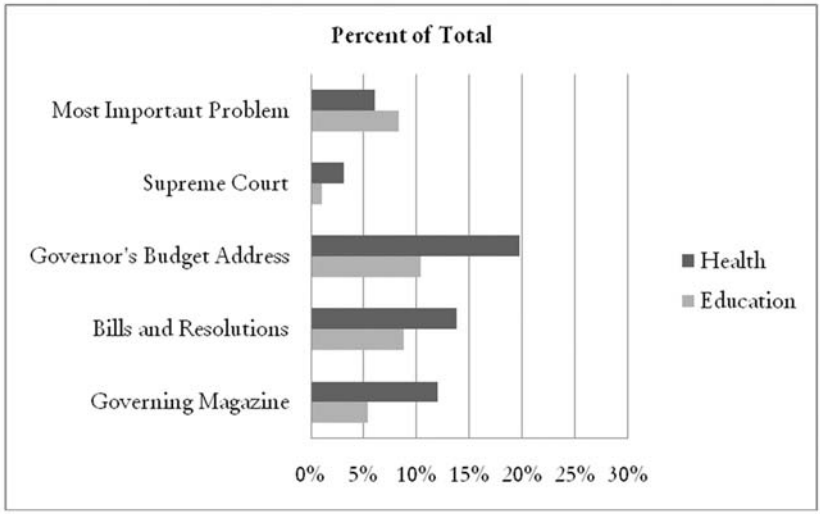


Table 2. Sample Records of Governor's Budget Message on Health Care in 2005

Day	Month	Year	Governor	Sentence	Source
9	2	2005	Rendell	While we succeeded in drawing down more federal funds for the Nursing Home Assessment and the Medicaid HMO Assessment, we continue to face even greater reductions in other areas of federal support.	2005–06 Budget Address
9	2	2005	Rendell	The joint federal/state Medicaid program effectively is the safety net for the poor and elderly and costs have risen dramatically over the last three years.	2005–06 Budget Address
9	2	2005	Rendell	Surging Costs for Medicaid Ravage State, Federal Budgets	2005–06 Budget Address
9	2	2005	Rendell	U.S. Medicaid Cutbacks Would Hurt States Twice	2005–06 Budget Address
9	2	2005	Rendell	Indeed, for the first time ever, combined state spending on Medicaid has outstripped state spending on K to 12 education.	2005–06 Budget Address
9	2	2005	Rendell	Medicaid spending is outstripping spending for education.	2005–06 Budget Address
9	2	2005	Rendell	We propose to limit the growth in drug costs by over 50 million dollars annually by establishing a preferred drug list for the Medicaid program.	2005–06 Budget Address

qualitative research design, one that is less susceptible to selection bias and that has already identified and incorporated important trends, insights, and turning points that might have otherwise remained hidden in mountains of immeasurable data.

Data Limitations

Ideally, all datasets would be either universal or, in the case of news reports, drawn from random samples. Although we believe we have captured the topics of all House committee hearings, and in many cases full transcripts, many Senate hearing records might not have been preserved. As noted above, state expenditure data could not be allocated to the minor topics and could not be fully allocated to major topics. Finally, partly because each record is coded only once (to enable users to measure relative attention across policy topics, years, and venues), coding decisions are contestable. The project manual provides database users with guidance as to where else to look for similar or related records.

ORGANIZING THE PROJECT: CHALLENGES AND OPPORTUNITIES

Organizing and executing a project of this scale presented numerous challenges to the project staff. This section summarizes four organizational and operational approaches that we believe can provide guidance for future projects.

Collaborations

The Pennsylvania Policy Database Project is a collaborative effort among the governmental and academic communities. After receiving financial support from the Pennsylvania General Assembly, we formed two advisory committees of state officials to seek input on the project and to keep state officials informed of our progress. The six-member General Assembly Advisory Committee is co-chaired by the secretary of the Senate and a special counsel to the House, and it also includes one senior staff member from each caucus. We meet with this committee at least once every six months to report on progress. A second committee included the directors of legislative service agencies and state records centers. This group served as an invaluable source of information at the beginning of the project when we were trying to map the location of government records. Our interviews showed that the records were dispersed throughout various archives and offices, but the cooperation of members of the committee allowed us to create a template for access and use of the records. Members of the committee also provided suggestions for

future archival and digital processing that we compiled into a report for the state legislature.

For practical and strategic reasons, Temple invited five other universities to join the project: Pennsylvania State University, University Park (where the faculty leader was Frank Baumgartner, who is now at the University of North Carolina at Chapel Hill); Pennsylvania State University, Harrisburg; Carnegie Mellon University's Heinz College of Policy and Management; the University of Pittsburgh's Graduate School of Public and International Affairs; and the University of Pennsylvania's Fels Institute.

We believed that sharing the work with universities across the state would broaden support for, and expertise available to, the project. We invited the faculty leaders from each campus to serve on a university advisory committee that contributed practical and academic advice on the project.

Each of the collaborating universities retained their own project managers and researchers. Although many datasets were available online and were therefore accessible to students on our Philadelphia campus, virtually all of the news digests were paper files located elsewhere in the state, which required students to work on location.[5] The decentralized nature of the project presented its own problems (discussed below), but we could not have completed the project without such collaboration.

Adapting Coding Decision Rules and the Codebook to the U.S. Model

The Pennsylvania Policy Database Project's codebook and decision rules are adapted from the U.S. Policy Agendas project. Except as noted below in the section on automated text coding, all records are read and coded independently by two student researchers, with graduate students or faculty members breaking ties. Following the U.S. codebook, each major topic is identified by a two-digit number (e.g., 03 for health care, 06 for education, and 08 for energy). Minor topics have two additional digits (e.g., 0601 for higher education, 0602 for elementary and secondary education, etc.). Each topic also has a minor general code ending in 00 for bills that involve multiple minor topics (e.g., 0600 for a bill dealing with higher and elementary and secondary education).[6]

The U.S. codebook emerged from a longitudinal examination of congressional hearings. Not surprisingly, these topics were not perfectly congruent with the activities of state government. Two state activities for which there are no federal parallels, and hence no explicit topics in the U.S. codebook, serve as examples: licensing of professional or commercial services and regulation of local governments.

The preferred option was to adapt the existing Policy Agendas Project

codes to state policy.[7] For example, on issues of licensing, we changed subtopic 0325 to Health Manpower and Training *and Licensing Issues* to account for state licensing practices for the medical professions. In a few cases, we created new minor codes. For example, under major topic 03 (Health), we created subtopic code 0345: Provision and Regulation of Ambulance Services. Under major topic 15 (Banking, Finance, and Domestic Commerce) we created the new subtopic code 1527 (Regulation of Services), as a catchall for topics that did not conform to other categories, such as licensing of barbers, beauticians, and undertakers.

State oversight of local governments proved to be the most difficult issue in adapting the U.S. codebook. Policy Agendas had a subcode for intergovernmental relations. However, given the importance and complexity of state oversight of local governments, we transformed a category used in the U.S. database only for news stories (24: State and Local Government Administration) into our 24, Local Governments and Governance. This topic has five subcodes related to the organization and regulation of sub-state governments, including home rule, changes in government structure, taxation, and debt. Substantive issues administered by local governments—such as education, policing, and recycling—are coded with their appropriate policy minor topic and are further identified by a "local government" filter. A researcher thus could call up all health bills that affected local governments.[8] We used the intergovernmental relations code for federal-state issues, preserving congruence with data from the Policy Agendas Project.

Our adaptation of the Policy Agendas Project's coding scheme to state government is based on our experiences with the operation of Pennsylvania government. Nevertheless, we believe that it is applicable across all 50 states. We consciously included subcodes for topics that were not germane to Pennsylvania (such as the initiative) in an effort to create a broadly applicable template for widespread use.

The adapted codebook is not a finished product. The creation of new subcodes and practical policy examples for subcodes is an iterative process. Difficult coding decisions were discussed among staff and researchers in an effort to find appropriate subcodes. In the instance of recurring "hard cases," we made changes to the codebook to provide guidance to coders. The primary difficulty was enforcing standardization across the collaborating universities involved in the project. We tried to mitigate this problem using a detailed codebook as well as frequent memos and PowerPoint tutorials to highlight difficult cases. Temple staff also made regular visits to the satellite campuses to refresh coders on the fundamentals of coding to ensure that individual universities did not create their own groupthink concerning the

coding process. All of the tutorials were then posted on our website so that project managers and coders could use them as necessary.

Records of Policymaker Attention to the News Media

The U.S. project uses *The New York Times* as a proxy for studying the media's influence on—or reflection of—the national public policy agenda. Although we initially tried replicating this approach, we quickly concluded that it was far too difficult and expensive a task and had conceptual weaknesses we could not overcome. We suspect similar projects in other states would reach the same conclusion. In the end, we created a dataset that measures government attention to the press (rather than press attention to the government).

No dominant newspaper systematically covers all of Pennsylvania, which is a large and complex state with five major media markets, the third largest rural population, and the third highest number of local governments in the nation. Sampling even two or three major metropolitan dailies would leave stories about policy issues in large parts of the state under-represented in the database.[9] Further, unlike *The New York Times*, no Pennsylvania newspaper publishes an index abstracting all of their news stories, which means student researchers had to read stories and create the abstracts in our database. Creating and coding more than 57,000 records in this dataset consumed almost 40 percent of the project's student hours.

Lacking a state newspaper of record, we abstracted and coded a 10 percent random sample of stories in news digests created on virtually every weekday of the year by state capital press offices and circulated to key state policymakers and advisors.[10] These digests typically contain between 15,000 and 20,000 articles a year, culled from large and small newspapers and electronic media across the state. Our sample of 1,500 to 2,000 abstracts a year, all of which are relevant to public policy, thus represents a universe of articles that are frequently reviewed by elected officials, cabinet officers, and staffers precisely because they reflect media coverage that they might not otherwise see.[11] In contrast, the U.S. Policy Agendas Project incorporates a one-percent sample (44,246 records) of all *New York Times* stories over 57 years and averages 781 abstracts per year, yet of these, only an average of 652 are about public policy, and the rest cover topics like weather forecasts, sports, and entertainment stories. As an example of the difference between the two datasets, the U.S. project can measure education stories as a percentage of all *New York Times* stories or all stories about public policy. Our database can measure education stories as a percentage of all public policy stories selected for inclusion in the news digests.[12] The policy topics in our dataset are thus more densely populated than those in the national database. Our abstracts are both a fuller outline of

newsworthy events and, we contend, a reasonable proxy measure of attention, albeit of the government to the press rather than the reverse. Scholars in states with a dominant newspaper might make a different decision, but their students would either have to abstract and code a random sample of each day's entire newspaper, which is hard to do without an index, or recognize and abstract only stories of importance to state policy. This is also hard to do because of the states' ubiquitous, but not always obvious, role in funding and/or regulating local governments, non-profits, and businesses. State press offices, we believe, will make better decisions about which stories are important.

Automatic Text Classification

Midway through the database construction process, we discovered, developed, and introduced computer-assisted coding, which has the potential to reduce costs substantially for our project and for similar projects in other states. Hillard, Purpura, and Wilkerson (2007a) report successful use of machine learning algorithms to classify congressional bills using the same codebook developed for the Policy Agendas Project.[13] Specifically, they reported agreement approaching 90 percent with the human coders for major topics and 80 percent for minor topics.

Applying the same algorithm (SVM) as documented in Purpura and Hillard (2006), we have been achieving agreement with double-blind human coders approaching 70 percent at the major topic code. When the major topic code is correct, the agreement at the subtopic code has been greater than 90 percent. Hillard, Purpura, and Wilkerson (2007b) have shown that agreement with human coders can be increased when multiple algorithms are used. They propose a strategy of accepting the computer classification when multiple algorithms agree and only apply human classification when the algorithms disagree. We have obtained similar results, but not to a degree that we trust the computer to be the only coder.

Prior to discovering the promise of computer coding, we had completed the double-blind human coding and tie-breaking process on all Pennsylvania Senate bills and resolutions and the House resolutions from 1979 to 2006. These roughly 34,000 bills and resolutions became the training set that we used to teach the computer to code approximately 41,000 House bills.[14] After the House bills were single-coded by humans, we applied the two best classifiers as described in Hillard, Purpura, and Wilkerson (2007b), and if they agreed with the human coder, we accepted the result. Human review and tie-breaking were used to resolve disagreement. This approach significantly reduced the effort required to complete the classification of the House bills.

Project researchers, working under the supervision of Professor Frank R. Baumgartner, applied a similar strategy to code news abstracts.

CONCLUSION

The Pennsylvania Policy Database Project responds to a number of the significant obstacles affecting research in state politics and policy. It provides easy access to a public website that summarizes a wide range of government and news media records systematically coded for policy impact from 1979 to the present. Because the project uses standardized decision rules and topic definitions and codes all records only once in mutually exclusive and exhaustive categories, it can quickly display trends within and across policy topics, decisionmaking venues, and historical periods. Because its policy codes also have been applied to the records of the U.S. government and are being applied to records of a number of foreign governments, it facilitates research in the fields of federalism and comparative politics. Although the Pennsylvania project's chief advantage is enabling users to recognize important patterns of policymaking across venues and over time, its embedded links also provide access to the full text of legislation and committee hearings, thus fulfilling the information-retrieval mission as well.

Temple will assist researchers who might wish similar databases in their states. As noted above, some of our datasets are already applicable to all 50 states: state expenditures, state general fund status, and *Governing* magazine articles. The Pennsylvania manual and codebook and our Access data entry forms could be easily adapted to other states, and we will share what we have learned from automated classification. Perhaps most important, we can share what we have learned from trial and error in project management. Construction of the Pennsylvania Policy Database Project has taken almost five years and will cost about $5 per record incorporated into the database. We believe that making early use of computerized coding and encouraging the state to scan news digests outside the parameters of the project would have allowed us to finish the job in two-three years at roughly half the cost. Costs for other states will vary according to the volume, location, and condition of records incorporated, as well as the rate of compensation for graduate and undergraduate students employed.

Robert Horton, Minnesota state archivist and director of *Preserving the Records of the E-Legislature*, a joint undertaking of the Library of Congress and the National Conference of State Legislatures, sees promise in the Pennsylvania Policy Database Project:

> Most state online systems replicate what you get in paper. They are notoriously hard to use. It is difficult to get an integrated picture . . . hard to connect the dots. You cannot justify investment in digitizing records unless you deal with public accessibility. I am very interested in what Pennsylvania is doing for that reason . . . . It is a different approach than I have seen anywhere else. (Personal communication with Joseph P. McLaughlin, Jr., May 24, 2008).

We believe it is an approach worth replicating across the states.

ENDNOTES

1. www.comparativeagendas.org (March 1, 2010).
2. www.congressionalbills.org (March 1, 2010).
3. Similar extensions of the U.S. Policy Agendas Project, all using the same coding system and decision rules, are under construction in Canada and a number of European nations. For more information and links to these projects, see the website for the Comparative Agendas Project directed by Frank Baumgartner at the University of North Carolina at Chapel Hill.
4. Nicholson-Crotty and Meier (2002) note the relative paucity of single-state case studies in journals and suggest it reflects professional bias toward 50-state studies. It might also reflect the difficulty of working in traditional state archives. A University of Massachusetts professor wrote that our project "is exactly what I wished had existed in my case study states. There was a paucity of print sources available, and it was very difficult to track down basic information about bill progression. I spent hours in each of these state's archives and came away frustrated . . . " (Personal communication with McLaughlin, July 27, 2006).
5. The news digests of Governor Dick Thornburgh (1979–87) were housed at the University of Pittsburgh, and those of Governor Robert P. Casey (1987–95) were housed at the University Park campus of Pennsylvania State University. Digests for the other years were archived in Harrisburg. In Governor Edward G. Rendell's second term (2006–10), his press office began sending electronic copies of the news digests to the project.
6. When a bill deals with multiple major topics (e.g., health care and education), it is coded by the dominant major topic. If it gives equal weight to more than one major topic (very few bills fall into this category), it is assigned the lower major topic code. Within the major topic for fiscal and economic policy (01), general appropriations bills funding the entire government have their own subtopic (0105).
7. Altogether, we made 38 changes to the U.S. codebook, including 22 minor topic additions and 13 minor topic changes, but the vast majority of state records are coded in categories that match or closely parallel the U.S. codebook. Titles of most minor topics were unchanged. However, we did update each minor topic to include relevant examples of state policy activity to assist coders. We did not eliminate any major or minor topics from our codebook, but we did identify 43 minor topics, such as international relations, for which there was little or no state activity.
8. The project uses a number of such filters to identify crosscutting features, such as taxes or budget impacts. Thus, a researcher could call up all health bills that include a tax, appropriate funds, or affect local governments, or combine all three features.

9. As an alternative to a newspaper of record, we attempted to recover all Associated Press stories generated by the Capitol news bureau for the entire period, but we were told by an AP executive that such recovery would be prohibitively expensive.

10. The project used the daily news digests compiled by governors' press offices, except for the administration of Governor Tom Ridge (1995–2002), wherein we used similar digests created by the House Democratic and Republican press offices.

11. In an in-depth study of the Illinois legislature, Susan Herbst (1998) concluded that legislators and their staffs regard press coverage of policy issues as the virtual equivalent of public opinion. Herbst writes, "All (legislative) staffers keep abreast of particular media reports in their own areas of committee work, via the internal clipping service within the legislature. These clips are very important because the staffers are not able to read all local dailies in Illinois, and they do need to get a sense of media coverage of issues across the state" (1998, 68). As communications director for the Pennsylvania House Democratic caucus in the mid-1970s, McLaughlin initiated the practice of circulating daily news digests to all caucus members. This process was quickly replicated by other caucuses and continues to the present.

12. Because the number of news stories in *The New York Times* index can vary due to fluctuations in advertising revenues and in state capital news digests due to fluctuations in collection efforts, both databases advise using percentages, rather than raw counts, to measure relative attention across policy topics and years.

13. Congressional Bills (www.congressionalbills.org), a sister project to the Policy Agendas Project, codes all legislation introduced from 1946 to 2000 using the same major and minor topics and decision rules. The Policy Agendas Project website classifies public laws (enacted bills) and roll calls on bills since 1946.

14. For computer coding to work, a strong set of previously classified data must serve as training and test data. This data must be consistently coded; otherwise, the computer algorithms will perform poorly. We used the computer to improve the quality of the training set in two ways. First, we used the computer to download and replace legislative abstracts that had been copied manually from the state website (www.legis.state.pa.us). This step eliminated human errors. Second, we developed an algorithm to identify bills with identical or nearly identical abstracts but with different assigned classifications. Contradictions were then resolved by human coders. This may explain why our initial results have not been as good as those experienced with the Congressional bills data. Other factors may include fewer instances of identical bills introduced by Pennsylvania state legislators and more variation in the drafting of short titles by Pennsylvania as opposed to Congressional reference bureaus.

REFERENCES

Baumgartner, Frank R., and Bryan D. Jones. 1993. *Agendas and Instability in American Politics*. Chicago, IL: University of Chicago Press.

Boyd, Don. 2005. *State Fiscal Outlooks from 2005 to 2013: Implications for Higher Education*. Boulder, CO: National Center for Higher Education Management Systems.

Herbst, Susan. 1998. *Reading Public Opinion How Political Actors View the Democratic Process.* Chicago, IL: University of Chicago Press.

Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2007a. "An Active Learning

Framework for Classifying Political Text." Presented at the annual meeting of the Midwest Political Science Association, April 14–17, Chicago, IL.

Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2007b. "Computer Assisted Topic Classification for Mixed Methods Social Research." *Journal of Information Technology & Politics* 4:31–46.

Kane, Thomas J., Peter R. Orszag, and David L. Gunter. 2003. *State Fiscal Constraints and Higher Education Funding: The Role of Medicaid and the Business Cycle* (Discussion Paper Number 11). Washington, DC: The Urban Institute.

Nicholson-Crotty, Sean, and Kenneth J. Meier. 2002. "Size Doesn't Matter: In Defense of Single-State Studies." *State Politics and Policy Quarterly* 2:411–22.

Purpura, Stephen, and Dustin Hillard. 2006. "Automated Classification of Congressional Legislation." Presented at the Seventh Annual International Conference on Digital Government Research, May 21–24, San Diego, CA.

Tandberg, David A. 2008. "The Politics of State Higher Education Funding." *Higher Education in Review* 5:1–36.