Civics in the Cloud Panel, Computing in the Cloud Conference
Center for Information Technology Policy, Princeton University
January 15, 2008
Joshua Tauberer, GovTrack.us

This panel, Civics in the Cloud, is, I think, aptly named for the three of us up here. It could have been named “Politics in the Cloud.” Then you might have expected to hear about how blogs, the long tail, and social networking are transforming electoral politics — campaigns. A hot topic. But while campaigns are about getting people *into* the government, “civics” seems to be more about the relationship between citizens and government itself. And that’s the focus for us, using the Cloud to illuminate and strengthen the relationship between citizens and government.

I am going to talk about three things. First, one way that computing today is helping to expand the relationship between citizens and the U.S. Congress by providing new views into the government. Second, what kind of policy decisions should be made so technology can do more. And third, where I hope technology will go in the future with respect to making the Cloud an integral part of civics.

The public’s primary Internet source for the status of pending federal legislation has been since 1995 a website named THOMAS, run by the Library of Congress. The site is today so important because Congress does far more than the mainstream media covers, or could cover, and if it weren’t for THOMAS the public could hardly learn about most legislative events. Between 5 and 10,*000* bills are introduced each year, and Congress only gets busier from year to year. With so much happening in Congress, we need computers to sort, search, and transform this information into manageable chunks for us humans.

A few things happened to pique my interests in civics and technology around seven years ago that eventually lead me to create GovTrack. Since this is my first return to Princeton in an academic capacity, I think I can take a moment to talk about how GovTrack began. The idea began a little after I took a class here as an undergraduate on the intersection of free speech, copyright law, and technology, taught by Professor Andrew Appel. As we were learning about the problems of the Digital Millennium Copyright Act and the seemingly unnecessary power Congress had given to digital content publishers, a story was simultaneously unfolding of Ed Felten’s watermark research getting him stuck between a rock — the recording industry — and a hard place — the law. The curious IT policy Congress had enacted made me intrigued by how little politicians are kept responsible to their constituents. I wanted to see how the Internet could spread information to bridge the citizen-congressman divide.

Eventually out of all of this came GovTrack.us, which went live in 2004. GovTrack is a tool you can use to track what’s happening in Congress. It collects information automatically from government websites, like THOMAS, and re-presents it in a number of new ways. One the one hand, GovTrack can be used as a research tool. One can search for and browse bills, much like THOMAS. Some other highlights of the site are... Color-coding of bill text to show how a bill has been amended on its way to becoming law. ... A Google Maps mash-up to help you find your congressional district. ... And some novel statistical analyses. The map here shows the political leanings of each representative, as computed by a statistical analysis of bill sponsorship patterns. You can also “Track” subjects, representatives, and bills, and get customized RSS feeds and email updates that are relevant for your interests.

This brings us back to our need for computers to transform information so we can see the same data from different perspectives, so we can understand it better. And for data about the government, it is especially important that the *public* be able to transform that data and not rely on the few presentations of the data that the government has the resources, mandate, or desire to create. GovTrack is of course just the beginning of what can be done with legislative information. Once you have a *database* of this information, there's no end to what you can do.

Well, almost. Sometimes databases can't capture reality when we don't know what the reality *is*, and that happens a little bit more than one would like. The Constitution defines how a bill can become a law. It outlines a finite state machine. But sometimes it's not clear when a transition has been made from one state to the next. There is a bit of a debacle over the defense spending bill**,** which President Bush claimed he pocket-vetoed in December. A pocket veto occurs when Congress is adjourned and the President neither signs nor vetoes a bill. A pocket veto does not allow for a congressional override, because Congress is not around to override it. This is important because in the case of the defense bill, an override would have been likely to succeed, which of course the President knew. However, congressional leaders claim they were just on a little break but not adjourned, and so no pocket veto could have occurred. They would view the situation differently: that the bill is not dead, but still awaiting either the signature of the president or a veto. There's a time-limit on that, though. The Constitution says that, if Congress *is* in session, after about 10 days of no action from the President, the bill just becomes law. To make matters worse, the President tried to cover all the bases by simultaneously going through some of the motions of a normal veto. So there are three states that the bill might be in: dead due to the pocket veto, or enacted into law by default after 10 days, or vetoed but awaiting an override.

Nuanced situations like this are hard to capture in data. It's one reason why the Congress is hesitant to publish raw data on its own, for fear of both publishing inaccurate information where it can't capture the nuances, and for fear of losing control of who is addressing the nuances.

But that doesn't stop us from building databases anyway. Other sites that use a legislative database include MAPLight.org, OpenCongress.org, WhereABill.org, and TheMiddleClass.org. In fact, these sites use the same database that I assemble for GovTrack. And that's because while the Library of Congress *has* a comprehensive database of legislative information, the one that powers the THOMAS website, it doesn't make that database available to the public in any machine-processable way, like XML, leaving sites like mine having to so-called screen-scrape the HTML pages of their website to reconstruct the underlying data. Screen-scraping is a delicate process and difficult to make comprehensive. It's a hack.

Machine-processable data, "structured data", is not a new concept in the government, but it is often resisted, and rarely made a priority. Some arms of the government have embraced it. The Census Bureau, the Securities and Exchange Commission, and the Federal Election Commission all make their gigabytes and gigabytes of information available not only on a website but also as a download in a machine-processeable format, and have done so for many years. Six state legislatures publish the status of their legislation in a structured data format. But while the web services team at the Library of Congress knows how useful a public database of their legislative information would be to the public, they had, last I checked, no plans to make their database available.

On the other hand, the House of Representatives is, actually, making some good progress with XML. They publish roll call votes and the text of most bills in XML. Putting votes in XML has made it easier not just for GovTrack but also the New York Times to create new visual representations of the outcomes of votes. Though..., the House's motivation has primarily been to improve their own internal work-flow, not really to share information with the public.

The Senate is catching up. They currently make no information publicly available using a structured data format. Besides the text of legislation and votes, the membership of committees is something they *have* in XML but simply don't share as a matter of policy. But the future is bright. At the December 11th hearing of the Senate's Homeland Security and Governmental Affairs Committee, Senator Lieberman, the chair, noted that Senate votes "are intentionally presented in a format that limits the public's ability to examine Senators' voting records," by which a staffer confirmed he meant that the Senate should use XML like the House does. Actually, this suggestion got to the Senator's attention through the work John Wonderlich and I, and others, have done with something called The Open House Project.

One of the problems with opening government data, especially in the Congress, is that goals for the use of technology are not being set. In The Open House Project, which John helped start and which is a project of the Sunlight Foundation, we outlined attainable technological goals for the House of Representatives, some of which I've just mentioned. Last month, an Open Government Working Group conference also developed a set of eight principles for what it means for government data to be "open". Some of the principles say the data must be made available in non-proprietary machine-processable formats, made available in a timely manner, and without special restrictions on its use. We offered the principles both as a benchmark and as policy recommendations.

So things are on the up and up for open government data here. We're getting more and more data that Congress already has opened up, in XML, and it is being put to use.

But I hope that XML is not the end game for government information. XML is a data format that here happens to facilitate proper standardization and normalization. XML is good for building a database. It is not good for meshing databases.

So corruption was once long ago a hot topic. Long ago being a year ago. Who was taking money from who, were they giving it back in earmarks, and were representatives voting with money in mind? To answer these questions comprehensively, three separate data sources are required: campaign contribution data, earmark data, and voting records. Databases exist for all of this information, but these databases don't talk to each other. They have no columns in common to establish connections from records in one database to records in another.

The direction I would like to see our civics-data community take is toward using the standards that make up the Semantic Web. The Semantic Web seems to be something that either you think is a hopeless dream of an early Internet pioneer, or else you just haven't heard about it yet. But I think it has a lot of practical value for civics.

The Semantic Web is a lot like today's web. Today's web is made up of web *pages* connected together by *hyperlinks*. Pages are something meant to be displayed by a web browser, for the immediate benefit of a human looking at pretty HTML and Flash. It's a web of *presentations*. The Semantic Web is a layer on top of the existing web whose target audience is computers, not humans. It is a web of *data*, and the point of it is to allow computers to help us process the *information* that's on the web. For example, Wikipedia contains a lot of information, but computers can't read the articles and make any sense of them. All they can do is present them on screen. A Semantic Web Wikipedia, which is being tried, would publish some of the same information but in a format that computers could manipulate, say answer questions about — though obviously not understand in any deep sense.

On the Semantic Web, websites publish databases in a common format called RDF, in such a way that records in one database can be tied to records in databases from other sources, if a certain amount of planning is done. The standards that make up the Semantic Web facilitate this goal. The benefit to civics is that we can start to ask new and more complex questions about our government, questions that span data sets, that we couldn't do before without a lot of hard work.

I've converted legislative, campaign finance, and census data into RDF, which means I can write queries in an SQL-like language that cover all of the data sets at once. How much money did lawyers contribute to representatives who voted for or against some particular bill? That query uses the sources highlighted on the projection. Did Rush Holt, Princeton's rep, receive more campaign contributions *per captia* from the zip code here, or from little-noticed sources elsewhere in the country? That needs all three data sets. It's not that these questions can't be answered today, but today's solution would be to create a new ad hoc program or database each time a new question arises. Using RDF, we make a trade off: a little bit more work in the beginning to create the cloud of data, but for the benefit of making each new query extremely easy to execute.

Some new questions lend themselves to new visualizations. The map I'm showing colors each congressional district in this area in red or blue according to the party of the representative from that district, in the usual way, but the intensity of the red and blue reflects the median income in each district. You can imagine there might be some policy debate for which this chart might be useful.

As more inter-linked data gets into RDF, the possibilities for meshing data together grow rapidly, which means questions about our government involving a lot of data become easier and easier to answer.

And that's the exciting thing the Cloud is doing for civics: making it easier to create new transformations of data, revealing new insights into what our government is doing, and in turn making the government more responsible to us citizens.

Thank you.