UPenn - Open Access Week at the Penn Libraries
October 23, 2013

Let me start with some of my favorite quotes about politics today. The first was about the debt negotiations between Congress and the President in 2011 right before the supercommittee was created. Remember how the Republicans "asked for too much" and President Obama "moved the goal post"? Reflecting on how the events were covered, reporter Matt Bai wrote for The New York Times:

> [A]fter the so-called grand bargain ... the two sides quickly settled into dueling, self-serving narratives of what transpired behind closed doors. ... [T]he whole debacle became the perfect metaphor for a city in which the two parties seem more and more to occupy not just opposing places on the political spectrum, but distinct realities altogether.

Politicians create a sort of manufactured reality this way. And it's been a problem for some time. In a classic 1988 essay called Insider Baseball in The New York Review of Books, Joan Didion described a scene from the 1988 presidential campaigns. On a hot day that year, candidate Michael Dukakis stepped off an airplane and had a short baseball toss with his press secretary. You can imagine how that could be an iconic moment in a campaign. It was widely reported on. Except there was no ball toss. At best we could say a ball was thrown. It was staged. And everyone knew it. But the journalists reported on it anyway. Didion wrote:

> What we had in the tarmac arrival with ball tossing, then, was an understanding: a repeated moment witnessed by many people, all of whom believed it to be a setup and yet most of whom believed that only an outsider, only someone too "naive" to know the rules of the game, would so describe it. . . . [T]his eerily contrived moment on the tarmac at San Diego could become, at least provisionally, history.'

And more recently I read this blog post in The New York Times about how campaigns are using data mining and other new techniques to get more votes, and again it's about the coverage of that by journalists:

> [J]ournalists remain unable to keep up with the machinations of modern campaigns. . . . Over the last decade . . . campaigns have modernized their techniques in such a way that nearly every member of the political press now lacks the specialized expertise to interpret what's going on. ... It's as if restaurant critics remained oblivious to a generation's worth of new chefs' tools and techniques and persisted in describing every dish that came out of the kitchen as either "grilled" or "broiled."

We're so naive, he's saying. And the politicians and campaigns get away with operating in these distinct realities because of information asymmetry. Government is big. There's no one person either inside or outside of government who knows how all of government works, or knows at least how the whole legislative process works, or even just knows how the House of Representatives works. Still, the people in power, inside of government, know government better than we do because at least they live it. And when we're trying to hold them accountable, we're at a huge disadvantage.

There are about 10,000 bills introduced in each two-year session of Congress. That's about 40 million words. It would take about 135 days straight of reading to read through them all, and then you would have to figure out what they mean. In each session there are thousands of committee meetings. It's amazing we have any idea what's going on.

So what are we gonna do?

If you watch Saturday Night Live you might remember this digital short. Andy Samberg's reaction to the system was to throw everything on the ground. I don't think we have to be that extreme. But I do think we, the public, have to stand up more and put in more effort to understand the system of government that we build.

Hacking is part of the answer.

What does hacking mean?

The answer to this question is pretty much in the next image. This is just some blog out there created by an IKEA shopping hobbyist. And the website is called IKEA Hackers. And in this post, Jules the IKEA Hacker, is showing how you can turn a pillow into a child's costume. This is hacking.

It has nothing to do with breaking into IKEA to get the pillow. The idea is to buy the pillow. And then, repurpose it. Turn it into something new. In a creative way. And solve a problem you have.

And this tweet from The Home Depot:

"Transform a door into a coffee table! The rustic upcycled design provides storage and is on wheels! Hashtag Home Depot Hacks."

It's hard to narrow down exactly what hack means. But I'll play linguist for a moment since I did study it here. As best as I see it, hacking is creatively solving a problem. Now there's this other word out there, it happens to be pronounced hack, the same way, that means something totally different. Hack is a homophone. Like gay. Remember the older meaning of gay. No one thinks homosexual individuals are happy, right? It's just a coincidence that we have these two words that happen to be pronounced and written the same way. But two meanings. Not related. I'm a hacker. But I'm not talking about cybercrime, leaks, or anything of that sort. And, in fact, my hack predates the use of hack to mean these other things. It goes back at least decades. I'm a civic hacker. I work on technological approaches to solving problems in our civic lives. Mostly about changing the balance in that information asymmetry that I mentioned before.

Here are some of the headlines for articles mentioning my work: Remixing government data, deep throat meets data mining, computer science in the service of democracy (that one was a little pretentious), and data mining meets city hall.

All right, let me tell you what I actually do. GovTrack.us is a free legislative tracking website. I hope you're as geeky about Congress as I am, otherwise that might sound pretty boring. We gather a LOT of data about Congress and help people understand what's going on, from what bills have been introduced to how your Member of Congress voted. It's a very data oriented website. Think of GovTrack as an app that teaches civics through current events.

I'll give some examples of what's on the website. Back in January when we all thought Congress might enact gun control, thousands of individuals came to GovTrack to see what these bills were really all about. Proponents wanted to know whether the bills were strong enough, opponents how much of the 2nd Amendment was being rolled back. Our users read these bills, checked which Members of Congress supported them, and many went on to track it --- which means we'll send them an email alert if a bill has any further action. That includes getting onto the House or Senate's schedule for the week ahead, getting a vote, gaining cosponsors, when the bill's text becomes available, and so on. Earlier in the year there was the vote on the Amash amendment to end the NSA's mass data collection. You can find that vote on GovTrack and see how the vote broke down. And you can use GovTrack to get an email every time Congress votes.

And if you're just getting started tracking Congress, you'll probably start by finding out who your representatives are. When you get to a page for a Member of Congress, you'll see a few of our own statistical analyses of Congress: ideology, leadership, and missed votes. This is our chart for missed votes over time. This one is for Senator Bob Menendez. Each point is the percent of missed votes, and it's over three-month periods starting when Menendez entered the Senate in 2006. So here in April to June of *last* year he missed about 2% of votes. We're also showing the context. Do you all think missing 2% of votes is good or bad? Does anyone want to guess the median missed vote rate in the senate? The bands are marking off percentiles. The first band at the bottom is the best 25% of senators, the second band takes you to the median which is actually right about 2%. So Menendez's record is right in the middle, not really either good or bad. We try to present context as much as possible.

Heading into last year's primaries, Representative Connie Mack took a beating in the press for missing votes while he was running for the senate. Mack deflected by saying his opponent, Senator Bill

Nelson, had missed even more votes. Politifact's Truth-o-meter looked up the numbers on GovTrack and rated Mack's claim true. So journalists come to us, and we also work with journalists a fair amount on pulling some custom data from what we have.

Do you remember how many bills I mentioned are introduced in every Congress? About 10,000. In the first 200 days of this year about 4,700 bills were introduced. Is that a lot or a little? What does that tell us about the gridlock on Capitol Hill? We did an analysis of historical data over the summer. And let me say, thank goodness we're not living in the '70s. Those congressmen wrote a ridiculous number of bills. In recent history, 4,700 is low. And be careful not to confuse that with a judgement about the size of government or the number of laws we have. Whether you want bigger government or smaller government, you still want Congress to be introducing bills to make that happen. It takes a new law to repeal an old law. So low doesn't mean less government, it means Congress just isn't doing much.

Do you know what percent of bills will be enacted? Let's say even before the gridlock today started. Let's go back to the Clinton and Bush years Let's do a show of hands. Everyone put your hand in the air. If you think Congress passes at least half of the bills it introduces, put your hand down. If you think Congress passes less than half of the bills it introduces, keep your hand up. Ok, so you all think it's fewer than half. Do you think they pass a third of the bills they introduce? If you think they pass a third, put your hand down. If you think it's less than a third keep your hand up. What about a quarter? What about just an eighth? A sixteenth? All right, well there were a few Congresses that passed as many as a sixteenth of the bills they introduced. But in most years it was about a twentieth or less, 4 or 5 percent. In 2011-2012, it was just 2%. Looking at the first 200 days of this year, we're on track for Congress enacting the fewest number of bills in modern history.

This is all really surprising to most people. Until you know how few bills are enacted, you think every story about a bill in Congress is important. You get the wrong idea about what is actionable information.

On GovTrack we give every bill a prognosis, it's a percent, the likelihood that we think the bill will be enacted. And like everything else, it's based on an objective statistical analysis, that takes into account factors that have, historically, made bills more or less likely to be enacted. Let's take an immigration bill by Rep. Chaffetz. We're giving this bill a 33% chance of being enacted. That's ten times more likely than most bills. This is a bill you should probably pay attention to.

This is context for understand whether this bill is relevant, whether as citizens and journalists this is something we should take action on and put energy into, or if it's something we can ignore. The prognosis is actually also a guide to why it might be important. Because we list the factors that go into the result. The sponsor of the bill, Chaffetz, is in the majority party in the House. That's crucial. He's also on a committee that the bill was referred to. That's really important. And he has cosponsors from the minority party, which is correlated with successful bills in the past. *And*, what's often not obvious at all, the prognosis tells us there's a history to this bill. Chaffetz introduced this bill last congress and it made it out of committee last Congress. That gives the bill a huge leg up this time around.

You'd know to look for all this if you worked in Congress. You'd know the factors and you'd have an idea of what those factors meant for a bill's chance of being enacted if you were on the inside. But here's this information asymmetry again. We don't know it, and that puts us at a disadvantage when we're trying to report on government. We have to guess what the factors are and then use statistical models to tell us which matter and which don't, and of those that matter whether they are good or bad for a bill, and how much. Each factor provides a little more context about what may be happening behind closed doors, and it might be a lead to an interesting story.

The prognosis is our most cited analysis. Often when a bill is *un*likely to be enacted. On the ever-colorful Huffington Post, they wrote that the Medicare Identity Theft Prevention Act has "only a 2 percent chance of becoming law. In other words, the Chicago Cubs have a better shot at winning the

World Series.”

The Washington Examiner last year ran a series of articles called “Tariff waivers are cash cows for Hill leaders.” They went through thousands of bills on GovTrack looking for MTBs, which apparently stands for Miscellaneous Tariff Bills. These bills create special exemptions from tariff duties. The Examiner’s staff compared the bills to MTB disclosure forms filed by congressmen to identify the companies that would benefit from these bills, and then looked for those company’s employees’s campaign contributions to see if the money was going full circle. It was a major research project. They found that Senators Bob Casey and Robert Menendez had introduced hundreds of these tariff break bills. Other Members of Congress, like Speaker John Boehner, had received hundreds of thousands of dollars from companies connected to MTBs. And, I think as a result of the Examiner’s analysis, there have been some efforts to reform this process, though I don’t know if anything has actually changed. If you’re interested in knowing which bills these were, the Examiner did a great thing and published their data spreadsheet to Google Docs.

The Washington Post did a similar story around the same time but instead of tariffs, they looked at the stock portfolio’s of Members of Congress. “More than 100 Members of Congress...traded stocks or bonds in companies lobbying on bills that passed through their committees.” One in eight trades by a Member of Congress was in a company they regulate. If it wasn’t insider trading, it was at least a conflict of interest. I think we provided the information on committee assignments, which was used to match representatives to the companies being investigated or regulated by a committee.

Everything I showed you so far is based on official, public information about Congress, and mostly even information that you can find elsewhere. But we try to make that information a) clear, b) instructive about how Congress works --- how it actually works not the middle school version of civics and not the whitewashed version of how government tells you how government works, and c) actionable like being able to track a bill. Then there’s how we do it. And to explain that, let me pose a question.

On a recent vote to reauthorize the Violence Against Women Act, representative Mike Rogers voted both for and against the bill. **Does anyone know how that’s possible?** And while you’re thinking about that, think of how important this question is if you want to do any data analysis about votes in Congress. There are two Mike Rogers’s in Congress. And this is what I worry about. Getting Congress to publish good data, like in XML like this, so that it identifies each Mike Rogers with a different identifier, and when Congress doesn’t do that sort of thing, we fill in the missing pieces so that we can do the analysis that we want to do. And all of that gets done in an automated way. So GovTrack is me, my webserver doing all of the data collection, and some part time staff to help with software development and social media.

So to sum up GovTrack: it’s data, which we repurpose into something understandable, into something analyzed and informative, and something shared. All of the data we collect is made available for anyone else to reuse and analyze. And GovTrack is also I guess a one-man policy shop. I nudge Congress and other government agencies to do the right thing.

Here’s what I mean on the policy side. If you go watch any video stream from the House of Representatives, you are first warned in scary quasi-legal language that “No portion of any recording may be used for a political purpose.” Well, what other possible use is there for government information than political purposes? And last year, a House committee said that they were concerned about budgeting money for confirming or invalidating third party analyses of legislative data. I’m that sort of third-party. They were concerned if they published more data, people like me would manipulate it and they would have to invalidate it. And then the Library of Congress –- a library! –- wrote in a letter to the House of Representatives that if they published more data they thought they might have an obligation to inform the public about the risks of the public actually using that data. Again, meaning websites like mine. It's offensive. What they've been saying is, the public can't be trusted with information about our government. We might confuse ourselves, or something.

Frankly, these are isolated cases. It's not usually this bad. Though, some times it's worse.

I'll wrap up with the case of the DC Code. That is, the legal code of the District of Columbia. Up until recently, if you weren't a lawyer with access to the pricey top of the line research tools from Westlaw and Lexis Nexis, your options for reading the law in DC were pretty limited. If you went online you would find a website like this. This was DC's official website for the DC Code at the time I took the screenshot. It was run by Westlaw. So... there was this guy named Tom MacWright who was looking for the local laws about bike lanes in the city. And what he found, in researching bike lanes, was that it was probably a felony for him to copy any of the laws about bike lanes into a blog post about it if he wanted to share what he found with others. And that was because there was a terms of service agreement that everyone implicitly agreed to when using this site that prevented you from copying any of the content. It read: "you will not reproduce, duplicate, copy, download, store, further transmit, disseminate, transfer, or otherwise exploit this website, or any portion hereof." And since that was the only electronic place you could find the DC Code, let's just say that effectively, Westlaw owned the law and could prevent people from telling anyone else what the law is.

Long story short, I worked with Tom and others on asking the DC Council for an electronic copy of the Code free of any copyright or other restrictions. I'm simplifying this story a lot, but about a week later they posted on the Internet a ZIP file containing 53 Word documents for the 53 titles of the Code. And they disclaimed any copyright stake in the files using a Creative Commons CC0 public domain dedication. So that was week 1. And in week 2 Tom built a whole new website for the DC Code. Here's Tom's site. It is beautiful, so much more intuitive, and its technology is laying the groundwork for a whole set of new tools that can help people read, disseminate, and understand the law. All of that in just about two weeks. In week 3, Tom asked a bunch of technology folks to meet up, at a hackathon, to play with the new data files to see what we could do. And who shows up but the general counsel for the DC Council, that is, DC's top lawyer in charge of publishing DC's laws. That's him on the left, dressed not like a top lawyer but like a coder, working with Tom on the right, at the hackathon.

And then I ramble about hackathons....