AzALL Legal Research Institute --- Congressional Information Symposium --- April 26, 2013

Joshua Tauberer

Keynote: Information access policy, consumer access constraints deriving from government policy, and on-line tools that break through policy barriers.

The narrative about politics today is that our government is stuck in some major gridlock. Partisanship is at an all-time high. The government risks being shut down. But, those are all lies. Political reality these days is entirely manufactured. Calling Congress gridlocked presupposes the parties are even negotiating, and it's just not clear anymore that that's even true.

Manufactured reality has been a problem for quite 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. A previous "ball toss" hadn't been captured well by the media, so the campaign tried it again. And, according to Didion, the journalists knew it was a staged event but they 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.'

The manufacturing of history is perverse and scary. And it happens regularly. Back in 2011, the public's view of the national debt negotiation between Congress and the President was a hodge podge of stories told by press secretaries. 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:

> [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.

And more recently I read this blog post in The New York Times about how campaigns are using data mining and another new techniques to get more votes:

> [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."

The problem that these stories highlight is about politics. They describe a world in which politicians are so far ahead of the public that they're able to construct whatever reality they want us to believe in. They can do this because politicians have a big advantage over the public: We often rely on them to explain to us what is happening in government. And politicians take advantage of this information asymmetry. But this problem is a part of a larger problem in government. Government is extremely complex. Just think of all of the time you've probably spent explaining how our government works to your library patrons. The complexity of government and the information asymmetry that creates are substantial barriers for all of us to be good citizens, or effective participants in our

government and legal system.

Any response to this problem ought to have two parts. There is a policy component to this problem. The more politicians turn government into a game, and the more complex government grows, the more information we need to effectively interact with government and keep it accountable. The second component, which is just as important, is: can we, the public, data mine as good as the politicians? This is our government, and we have a responsibility as citizens and government professionals, too, to be engaged even if our government doesn't make it easy.

And that's where I'm going to start.

* * *

About 13 years ago I started hacking the law. I saw that the U.S. Congress puts out a lot of information. And at the time, there weren't many good ways for every-day people to make any use of that information. There was no way to know what was important, or what anything meant. I could find out how my representative voted on Roll Call Vote 53, but what did that really mean? It was just a part of the information overload that creates information asymmetry. I thought, at the time, that if we had better tools to track Congress we could keep them more accountable. I looked for as much data on Congress as I could find so that I could build free tools to help people navigate the information coming out of Congress. I launched GovTrack.us in 2004 to do this. Today GovTrack is one of the most visited government transparency websites in the world, and it educates mores citizens about what is happening in Congress than Congress's own websites THOMAS and the new Congress.gov.

There are two parts to GovTrack. One is what you see when you visit the website. You can find the status of every bill in Congress and get updates as bills move through the legislative process by email or RSS. You can read the text of bills. The site has bill text and status back to 1993, title and status back to 1973, and information on bills that eventually became statutes going back to 1951. You can find your representative with detailed congressional district maps that you can zoom in on to street level, and you can get updates about what your representative is doing. Or track votes as they happen, almost in real time, or go back in history all the way to 1789 and the first vote in our nation. For bills, the advanced search can filter by Congress, by the current status of the bill, by sponsor and all simultaneously. Of all of the features I'm mentioning, this is one of the most important for research. You can search bills by slip law number. And then there's automatic red-lining, which is one of my favorite features on the site. It shows how the text of a bill changed as it goes through the lawmaking process, or how it's changed from Congress to Congress if the bill has been reintroduced.

The site also has some unique statistical analyses of what's going on in Congress. Our ideology analysis assigns a liberal–conservative score to each Member of Congress based on his or her pattern of cosponsorship of bills. In this chart here, the names are current senators. The left-right position of each dot is the ideology score. In a nutshell, Members of Congress who cosponsor similar sets of bills will get scores close together, while Members of Congress who sponsor different sets of bills will have scores far apart. The analysis is totally blind to what any of these bills are about, or political party affiliations, but it's able to infer something real about the politics in Congress. Look at McCain but don't worry about Flake. I also compute a leadership score for each Member of Congress. That's how the dots are positioned on the up-down axis. The leadership score is computed by looking at who's scratching whose back. The idea behind the leadership score is that if Senate X cosponsors Senator Y's bills but Senator Y does not cosponsor Senator X's bills, then X is a follower relative to Y being a leader. The analysis uses the same algorithm that Google uses to rank pages on the web, which determines the order of search results that you see when you do a search. Again, McCain and Flake.

And GovTrack computes a prognosis for each bill. That is, for any bill, what is the chance it's going to be enacted. More than 10,000 bills will be considered by each Congress. About 3% will become law. Which bills should we focus on? The prognosis tells you which bills you can probably ignore. When you're worried Congress is going to appoint President Obama a dictator for life --- and that's a real preoccupation of some of my users --- it's helpful to know that the resolution to take away

presidential term limits has a prognosis of 0%. If you're a lawyer or government affairs professional who wants to know how the law might change soon, the prognosis will help you manage the information overload too. Here's the prognosis for H.R. 1123: Unlocking Consumer Choice and Wireless Competition Act, a bill by Republican Representative Bob Goodlatte. This bill is probably going somewhere with a 90% chance of being enacted.

But the prognosis not only gives a number, like 0% or 90%, but it also highlights the factors that make a bill successful. Such as, is the sponsor the chairman of a committee? Is the bill naming a post office?

These scores, the ideology, leadership, and prognosis, give us a view into Congress that is normally hidden to us. We can't observe leadership, not directly. We're not there, in Congress, to see it. We're not in the meetings where you can see relationships form. But those relationships are known to the congressmen and senators. It's obvious to them. They know whether they lead or follow. Their staff know. This is a sort of social knowledge that is locked within the institution of Congress, unless we get a little creative with how we try to observe it. And congressmen and senators, their staff, and lobbyists all know what bills are important because they have in their brains the institutional knowledge of what makes a bill important like some of those factors that I mentioned before. On the outside, we want to know too, and data mining can help us do that.

I talk about GovTrack often from the citizen's perspective. About 75% of the site's users are everyday citizens. They're looking up who represents them in Congress. They're tracking votes and reading bills, even reading the text of bills. And probably most often they're researching a bill they heard about through an advocacy organization that they are a member of.

But I hope you see the practical value for students, researchers, and legal professionals too, for how technology helps us make heads or tails of the huge amount of very complex information coming out of our government. About 15% of GovTrack users are journalists or legislative professionals. Many of those are legislative affairs people for small businesses that can't afford or don't need expensive legislative tracking tools like "Wexis." Half of those people are in the DC metro area, so they're probably lobbyists and other policy staff. And half of those, or 3% of the site's visitors overall, are staff in the House and Senate.

GovTrack is an example of an open government application. There are lots more, and I know you'll hear about a bunch throughout the day. The open government movement has exploded in the last few years, and so now you'll find apps that help you navigate government and law, like GovTrack, all the way to apps that use government data to rate financial advisors. It's not all about civic engagement.

* * *

So as I mentioned, there are two parts to GovTrack. And this applies to most open government projects. The first part was what you see. The second part is getting all of this information. This part is about data. What a lot of open government apps have in common is we get existing data, or at least existing information from official government websites. Collecting new data is really expensive. We civic hackers look for what we can do cheaply. In the best case, the government is publishing structured data for what you're interested in. Structured data is the key component of the modern open government data movement. It's our gold. Programmers like me, and Eric Mill who's speaking later, and researchers, and designers, and librarians, we can take this structured data and transform it into tools. Like GovTrack. Like the weather maps you see on TV --- they come from government data. Online road maps, like Google Maps, were at one time based on government data. Even FederalRegister.gov: the official site for the Federal Register was originally built as a private-sector website using public structured data. It was so good it became the official website.

So what is this structured data? Let me explain by analogy.

I really don't get poetry. I can read the words, I get the rhythm. I can even appreciate the visual aesthetics. But I don't know what it means! Middle school English was a nightmare for me. And this explains today why I really like rap music, because I can't understand the violent things the rappers are

talking about through the poetic ways they say it. *I* only have trouble with poetry. But computers have trouble with everything. Prose is like poetry to a computer. Point a computer at a Wikipedia page and it's got nothing. Computer can't understand English, etc.

Structure is what makes information processable by computers. Spreadsheets, data standards, and other sorts of structure makes it possible to reliably extract information out of data. A little structure goes a long way toward making this document analyzable and reusable.

The Senate and House publish roll call vote results in a structured data format called XML. Here's what it looks like. Each Member of Congress is identified by a unique ID so we can reliably know which Member of Congress corresponds to which vote. This is from the Violence Against Women Act reauthorization vote, and you can see that by using IDs it's possible to see which representative voted in favor and which voted against. It's a lot better than names. Because in this vote, Rep. Mike Rogers voted in favor. And Rep. Mike Rogers voted against. There are two. But one is R000572 and the other is R000575. Now the important point is that I'm not saying it's impossible to figure out who is who without that ID. If I had to, I can just have an intern call up their offices and ask them how they voted. What structure does is it makes information processing so much cheaper that it makes entirely new things possible, like the sorts of things I've done on GovTrack.

That was votes. We have structured data for votes. But there is no corresponding structured data for the status of bills. There's a website, actually now two websites, from Congress that list the status of legislation. THOMAS.gov, which dates back to 1995, and Congress.gov, which launched seven months ago, provide tons of information on legislation. But there's no data, so I have to scratch my ear like this <!> to basically reverse-engineer what's in THOMAS's database. I programmed GovTrack to crawl THOMAS each day scouring for new information like new bills and new activity on those bills. I can talk more about this process during questions if there's interest, but suffice it to say, the process is fragile. I might be missing a bill on GovTrack, or the title might be out of date, or sometimes, rarely information is wrong. Without structured data, recreating this information is hard to do accurately. If Congress shared the database of this information they already have, there would be no issue of accuracy because I would have the same data Congress has.

You all know refrigerator poetry? Let me push this metaphor even further. THOMAS.gov is like poetry. But it's like refrigerator poetry that's been glued together into one, permanent haiku. I want to rearrange the pieces and build something new, and different, and interesting. You can appreciate the official haiku, sure, but the glue has limited the potential that comes from being able to rearrange the pieces and discover new meaning. Letting the public have access to the legislative status database that Congress already has is an easy win for government information dissemination and public education.

I asked the Library of Congress if I could write my own haiku, so to speak, ten years ago. That is, I asked if I could have their database. Here's their response: "Use THOMAS.gov // that should be enough for you // we don't share data." I'm paraphrasing here. It's a haiku. The policy, coming out of the House and Senate committees that oversee the Library of Congress, was that Congress should control the way in which this government information is presented. Not only don't we, the public, have a right to the data, but the prevailing wisdom has been that we might confuse ourselves with it. So, no data for us.

The irony, of course, is that the public is woefully confused already. It's abundantly clear from the last five to ten years that more data makes everyone more informed, not less.

* * *

The THOMAS database is just one example of what's going on in what's now a pretty complicated landscape of the open government movement. I mentioned this movement has exploded. Historically, starting around the 1950s, the open government movement was what we call today the right-to-know or the freedom-of-information movement. It was based on the idea, promoted by journalists, that there is a legal right to information held by the government. And it was a disruptive change in 1966 when FOIA --- that's the Freedom of Information Act --- was enacted.

The legal right creates a presumption of openness, but, as you know if you're familiar with FOIA, the right is not pro-active, it's reactive. If there's data you want, and you can figure out which agency has it, you can basically petition for that information. And if you're lucky, the agency won't object and claim one of the exemptions, if you're lucky the agency won't make you pay much to have the data retrieved and copied, and if you're lucky you'll get it in about a year. Today, almost 50 years after FOIA was enacted, it's pretty obvious we can do a lot better than that.

You should all have the open government data maturity model handout. This is a road-map I am proposing for how to implement open data. This is what I want Congress, or any government body, to do. Down the rows on the left side are the different technological strategies of OGD: freedom of information, using the Internet, principles of openness, structured data, global IDs, APIs, and linked data (the semantic web). Across the columns at the top are the different sorts of public information governments produce: laws, service-related data, data about the structure of government, operational data such as rulemaking dockets and spending records, and finally a catch-all column for other public data (sometimes produced incidentally to government functions). The rows and the columns have an order to them, which makes this a maturity model.

You can use this as a bingo card as I talk through what it means. Check off the boxes that I say Congress has done!

The rows of the maturity model are ordered according to their technological complexity. Each successive technology makes data more useful for searching, sorting, and transforming the data to new purposes. The order is not from cheap to expensive. In fact, it may be just the opposite. For instance, the total cost of all FOIA-related activities across the federal government in FY 2008 was $338 million, mostly for the 3,691 full-time-equivalent staff processing FOIA requests. The first row of the table does not come cheap. Technology helps us reduce costs in the long-term. Each row in the table let's us do more with less.

So the first row is FOI, a basic legal right to documents held by the government. As you may know, the federal Freedom of Information Act applies to just one of the three branches of government. So right out of the gate you can see the legislative branch has a long way to go on this road map.

The next step says, make FOI pro-active rather than reactive. Instead of waiting for a FOIA request, and then possibly mailing a CD-ROM, governments should pro-actively post frequently requested documents to their websites. Because while FOI might satisfy a very narrow sense of public access to government information, information is not meaningfully public if it does not reach a wide audience. Public information must be online, accessible without paying a fee, and it has to be findable. Since the launch of GPO's FDSys website in 2009, it's become a lot easier to find congressional publications online. Still missing: committee votes.

Now that I have the document, the next question is whether I am able to use it. There are a few components to this but let me give one example from legislative video. If you are watching a video of a committee hearing on the Library of Congress's website, you are told: "No portion of any recording may be used for a political purpose; no portion of a recording may be disseminated with commercial sponsorship except as part of a bona fide news program or public affairs documentary; no portion of a recording may be used in any commercial advertisement; and any redistribution must be subject to this same notice." Gosh. Let me rephrase this. Our government is telling us, we'll show you what your representatives are doing in public meetings but only if you promise not to use the information to educate voters about how to vote, which sounds like it would be one of those political purposes they're talking about. And, if I redistribute it, well, GovTrack runs advertising to stay fiscally self-sufficient, so I guess I can't disseminate that government information. Yikes.This is a back-door offense to the first amendment. One of the core principles of our government is that they don't censor political speech. And so a core principle of open data is that you don't have to agree to a contract, a terms of use policy, or a copyright license in order to get or use the data.

The next row is “Structured Data”. This is the first row that is purely technical. It means, use spreadsheets instead of PDFs, use text instead of scanned images. And make the data available in bulk, as big downloads, as much as possible. Or use XML. Break down fields into processable components. Most importantly, make the data useful to people who can search, sort, transform, process, and analyze it. There has been some recent progress in the use of structured data in Congress. I mentioned voting records in XML. But as I mentioned, we’re still missing some of the most important information. Like, a list of bills introduced in Congress with their titles. Even something as basic as that Congress hasn’t gotten behind.

This is as far as most government agencies have made it on the technology of open government data. The remaining three rows guide future directions. One is to assign IDs to things. The R000572 and R000575 IDs for the two Mike Rogers’s are a step in this direction.

What are APIs and linked data? I don’t know of any work on APIs or linked data in the legislative branch, but I’m glad I don’t. The legislative branch needs to learn to walk before it learns to run. We’ll get to linked data maybe in a few years.

The columns at the top of the maturity model cover the different sorts of public information governments produce. The columns are in a particular order from left to right. Whereas the order of the rows is based on a logical technological progression, the order of the columns is based on a set of normative values relating to the purpose of government. Reasonable people may disagree on this order.

The columns start on the left, where there is a moral imperative for the government data to be made available to the public, and end on the right, where access to public data creates additional benefit to society but for which there is no moral imperative to make the data available.

The leftmost column is “Law”, and here the maturity model asserts that access to the law is the most important function of the many purposes OGD serves. A moral imperative to promulgate the law in all of the ways that increase access stems from the principle that ignorance of the law is never a defense. The principle is quite a conundrum when the law is hard to find, difficult to understand, and, at times, illegal to share (though that’s only a problem at the state and local levels). The moral imperative is only a starting point. Access to law has wider implications from improved civics and law education in schools to reduced costs of legal compliance for small businesses. At the federal level, the United States Code and the Code of Federal Regulations are being published now as bulk HTML or XML data. Two for three branches isn’t so bad!

Services are next. I’ve revised the order after making the handout. Services are data produced in the furtherance of a government program. Weather data is an example. The National Weather Service is, or at least was at one time, the largest producer of public data in the government. The Census was one of the first agencies to put data on the web. Their data is another example of service data. For services it’s not a moral imperative but a legal imperative. If an agency’s mission is to produce information, publishing that information as open data can help it do that better. Congress doesn’t really provide services to the public, except perhaps the service of reading and replying to constituent mail. The House has been working on an API that would receive constituent mail electronically and send it off to the right office. Kind of like email. It would mean that third-parties could more easily help constituents get their letters into Congress by building tools that interact with the API.

The middle columns are Structure and Process and Spending. This sort of data is information about how government is being run and how money is being spent. This is where government accountability looks for corruption. There is a moral imperative here too, rooted in the idea that only an educated public can hold their government accountable. For legislative data, this would include a list of Members of Congress and their committee assignments, the text and status of bills, voting records, Congress’s own legislative-branch spending records, and so on. All of this information goes toward making citizens better participants in our own government, which was one of the problems I

highlighted at the beginning. I've already mentioned that we have some of this in structured data and some not.

And finally the catch-all column for public data. This is, for instance, some sorts of Medicare and Medicaid claim statistics. Or geographic data about the location of every single road in the country. There is some data like this that there's no moral imperative to make public, and there's no legal imperative to pro-actively make it available. In a resource-limited world, this sort of data is not as a high priority for open data as some of the other sorts. But making the data open, structured, and so on produces value to society. It's civic capital. Entrepreneurs can build businesses around this data. Like Google Maps. Think of the all of the time online maps has saved us in our lives and the value of that. A lot of those maps started as government data. Congress produces data like this. It has a think tank called the Congressional Research Service which writes a huge number of reports for Congress, but those reports are not made publicly available. Again, no moral imperative per se, but those reports would be hugely educational for the public. They have huge value. They should be made available.

So now I'll wrap up. Here's everything I said in a nutshell: data mining our government is the only way for us to learn about how our government works and how we can participate in it. Our government is complex. And that's true whether you're a citizen, a lawyer, or a librarian. Digital tools that transform raw government information into something you can learn from and use has enormous power for empowering participation, educating the public, and changing the citizen experience. But we need a strong and well thought out legal policy to guide and nurture the development of open government data.

Plug my book!

[47:00]

ADDS

My classmates at Princeton were building new music sharing tools on top of the campus's network infrastructure. My professors were conducting research on digital watermarks, and finding that watermarks were not very good at stopping file sharing. Things were looking up for free expression and creativity. Until my professors were threatened by the recording industry not to publish their unfavorable research results. One of my classmates was among the first four students ever to be sued by the recording industry for peer-to-peer copyright infringement. All the while Napster and other peer-to-peer networks were being sued out of existence. The recording industry's response to music sharing was harsh, felt harsh, at Princeton it was personal, and it motivated many of us there to use our technology expertise in the public policy sphere.

I was new to politics then. I hadn't yet ever voted. But I saw what I thought was an untapped resource.

One other example I want to mention right now is the case of the 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 West and Lexis, 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. 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 Code 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 the 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, West owned the Code 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 the West site again. And now 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. This is one of the most successful and rapid open government success stories that I’ve ever seen.

We’re a long way from DC, I know, but the principles are the same everywhere. Information technology applied to the legislative process, from bills to statutes to codes, can enormously reduce information asymmetry.