Legislative Data Process Conference
February 2, 2012

Data Impact / Data Quality

You all know what refrigerator poetry is right? The little magnetic pieces that you put on your refrigerator and you make poems by rearranging them. I was once at a store buying refrigerator poetry like you do, and I got home and opened the package, but to my dismay, all of the pieces were already glued together. It was a pre-made sonnet. The sonnet was great. It really touched me. But, I wanted to make one of the other 40,319 poems that could have been made with those eight pieces. The value of refrigerator poetry is being able to rearrange these bits and come up with something new. Data is a lot like refrigerator poetry. You can come up with all of these things, some of which will be interesting, some of which will engage the public and really make the world better. Congress has a lot of information and makes a lot of data available to the public so it is in a great position to encourage the public to innovate and promote civic engagement.

I want to talk about the impact side of this a bit. I lauched GovTrack.us in 2004 and it's a bit like THOMAS with bill tracking, bill search, reading bill text, and here I'm showing Congressional district maps, a Google Maps mashup using Census GIS data files. Some of these features are on THOMAS. Some are not. Many are not with the scope of the Library of Congress's mandate to provide. Maybe it's not up to Congress to make these maps. Not all civic tools can be something Congress could or should provide. But counting over the last six months, 5 to 10 million individuals have used GovTrack, or a website that reuses the data that GovTrack puts together. That's a lot of people that are being helped and engaged. Now, even though I built a lot of those tools, the Library of Congress can *also* take credit for it because so much of the information comes from the Library originally. Lemme repeat that: The Library of Congress and Congress should take credit for this work.

Sometimes Congressional data goes out to help the public. Sometimes it comes back to help Congress do its job better. At POPVOX, we're building tools to help constituents be better advocates for their issues. And that does *not* mean building more antagonistic advocates. We've helped constituents write around 400,000 letters to Congress over the last year, and even though the problem that we're trying to solve has a lot to do with the fact that Congress is overloaded with constituent mail, legislative correspondents consistently tell us they wished everyone were writing through POPVOX and that's because we know how to get the messsages in in the way that makes it easier for LCs to process that information. Solving the problem of constituent correspondence makes constituents better advocactes at the very same time as helping the work flow in Congressional offices be more efficient. POPVOX wouldn't be possible without the legislative data coming from the Library. We need that large database of information in order to give constituents the right user experience so that we can help them be better advocates which in turn helps Congress process its mail better.

There are other ways open data helps internal workflow here. I've been working on change tracking for bills for many years. If you're looking at a long appropriations bill like H.R. 2112 and you've been following it through the process. Maybe it's engrossed now. And you want to figure out what's in the bill now that it's been passed. It's really helpful to be able to see the changes that have been made from version to version as the GPO prints it, rather than having to read through thousands of pages after each print. What I've been able to do is create something like a markup that shows the changes to each version with red underlining right on the PDF itself. It is a layer of machine processing added on top of the bill that helps you understand what is in the bill.

There are other things I wanted to tell you about how open legislative data can help internal processes here, but I'm told I can't because it would be promoting a product we charge for.

Lots of things come out of data. Some things are useful, some are not, some you don't know. That's okay. You have to cast a wide net and see what comes out of it. Dan Nguyen did an interesting

analysis of the smiles of Members of Congress. He analyzed photos of every Member of Congress and measured how large each's smile was. And I don't know if this is useful or not but it is certainly thought provoking.

Sunlight Foundation did a fantastic project called Capitol Greeting Cards. They're like e-cards you can send your friends during the holidays. If you find this card on Sunlight's website and click on the card, you'll hear this representative Mark Kennedy from 2003 quoting from Star Wars on the floor of the House. This is a combination of congressional information and artistic skill. The subject matter and the presentation together is very engaging I think, and that's really important. Obviously the Library of Congress's job is not to make greeting cards that are a little tongue in cheek, but being able to create these is really important for civic engagement. Congress should be promoting civic engagement by making the data available that makes it easier to build these sorts of things.

So I've talked about impact. I've talked about millions of individuals being helped by legislative data. I've talked about Congress being helped by legislative data. And I've talked about how data and art is important for civic engagement and must be carried out in the private sector.

Now I want to talk about the data.

For GovTrack and POPVOX we use a lot of Congressional information sources, including THOMAS, the House and Senate voting records, GPO's FDSys, the Census's GIS data, and cost estimates from CBO, among others. Some of these databases are done pretty well. Bill text is published pretty well. Senate committee schedules pretty well.

But some things are not done well. There's no data coming out of Congress on comprehesive bill status. There's a website, THOMAS, but there's no database. So I scratch my ear kinda like this to round up all of that information and get it entered into a database, as cheaply as possible frankly. That's what happens when Congress doesn't provide its own good data. I end up doing it. You know what else happens when Congress doesn't do it? People think I'm the authoritative source for legislative information! If Congress wants to make me the authoritative source for their information, that's great for my businesses. I can make a ton of money doing that. But I'd be so much happier if Congress could be the authority for its own bill status information.

The question is here, what am I actually asking. It's a big ask. It's so big that I'm writing a book about it at opengovdata.io --- dot i o. There are 17 different principles of open government data including machine processability, which is the big one, licensing, digital signatures, bulk data, and other principles. Also I've put up a case study on the House disbursements data release, which does really well in a lot of categories, especially especially documentation. The disbursements website has the best documentation that I've seen for any government data source. But it punts on some of the other important aspects of open government data, like machine processability, which, again, is the most important one.

How do you know the disbursements PDFs are not machine processable? Or that THOMAS is not an effective data source? You need a concept of data quality. Data quality is, to me, the intersection of accuracy, precision, and cost. Precision is the depth of knowledge encoded in the data, sort of the amount of information there is. Accuracy is the likelihood that if I try to tell you what's encoded in the file that I'd actually get it right. Voting records coming out of Congress have high precision and high accuracy. The use of alphanumeric IDs to identity Members of Congress and an XML format means that it has deep, precise encoding of what happened in the vote, and also it is accurate because it is reliable. I can tell how everyone voted accurately by reading the files. House Disbursements has high precision. It gets down almost to line items. I think that's great. But it has extremely low accuracy because it was published in a PDF, which makes it very difficult to reliably put the information into a spreadsheet so that you can do any sort of data analysis. It's possible to do it reliably, and that's where cost comes in. So precision, accuracy, and cost --- I'm hoping those are useful terms for framing conversations going format.

What's government data good for?
Civic Capital (maps, weather, etc.)
Transparency & Accountability
Democratizing Access to the Law
Citizen Engagement with Government
Reducing Cost of Legal Compliance for Small Businesses
Enhancing Educational Materials

Data quality examples:
bill text: accurate but not precise
house committee schedules: neither accurate nor precise
amendment text: neither accurate nor precise
US Code: with respect to sectioning, precise but not accurate
what of cost? hiring lots of interns ok?