Wolfram Data Summit
September 9, 2010
Joshua Tauberer

Perspectives on Open Government Data Policy

Earlier this year I was asked to give a talk on designing government transparency, at the Center for Information Technology Policy at Princeton. Let's say we could start from scratch, what would a technologically-enabled open government look like? Would it involve mandatory open records? Live streaming of all public meetings? Maybe the right to information would be a constitutional right:

"(1) Every citizen has the right of access to information held by the State. (2) Every person has the right to the correction or deletion of untrue or misleading information that affects the person."

In fact, this is no hypothetical right. I took this from the new constitution of Kenya[1] which was ratified last month. But Kenya wasn't the first to make information a right. According to the website Right2Info.org[2], the right to government held information is protected by the constitutions of around 40 countries. And although the United States is conspicuously missing from this list, access to government information isn't a new idea. Apparently the Swedish constitution has had a right-to-information provision since 1766, back when the country was transitioning from a monarchy to a constitutional monarchy. Their provision today reads: "Every Swedish citizen shall be entitled to have free access to official documents, in order to encourage the free exchange of opinion and the availability of comprehensive information." The Philippines has a particularly strong provision as well specifically mentioning records, documents, papers, and *research data.* Here's the full list from Right2Info.org, minus a few countries that I thought weren't strong enough to include here.

So anyway, what I said at the conference back in Princeton was that government transparency is a paradox.

I'm kidding. A little.

Since 2004 I've been running GovTrack.us, which is a website that tracks the activities of the U.S. Congress. It's also a platform that others can use to track the U.S. Congress. GovTrack's value-add is bringing congressional information into one place, in a unified data model, and making it available to users for free. Most of the information on the site can be found elsewhere, but in so many difference places and in formats that are barely useful to the American public. For instance, voting records are found for the House of Representatives on the House's website and for the Senate on the Senate's website. Recently the Senate joined the House in publishing voting records in XML, but the two bodies of Congress use widely divergent XML structures. The status of legislation is available from the Library of Congress. And so on. So through lots of Perl scripts GovTrack screen-scrapes all of these websites, normalizes the information, and creates a large database of Congressional information containing: bill status, voting records, bill text, biographical information about Members of Congress, committee assignments, and some more.

Here's the site. GovTrack was one of the first websites world-wide to do this sort of thing with government data, and is probably the only one that runs a profit. It reaches about half a million people each month directly, and well over a million if you count visitors to third-parties websites and mobile apps that are based on GovTrack in one way or another.

1 http://www.nation.co.ke/blob/view/-/913208/data/157983/-/l8do0kz/-/published+draft.pdf
and for more: http://j.mp/98AtiE

2 http://right2info.org/constitutional-protections-of-the-right-to#

When I opened up the source *data* that powers GovTrack in XML and RDF formats, others started to see the potential for building other tools that shed light on government processes in new ways. The two biggest reusers of the data are OpenCongress.org, whose value-add is a blog that covers the activities of Congress in a way that's clear for those of us outside the beltway, and MAPLight.org, which puts a new spin on the connections between money and politics. Another interesting one is IBM ManyBills, which is a visualization tool like their ManyEyes but for congressional legislation, and there's a mobile app called Congress built by Eric Mill at the Sunlight Foundation. At least two dozen websites have popped up relying on GovTrack data all trying to give the public a new way to get a grasp of their government --- I'm sure there are a number I'm not even aware of.

Structured, normalized information about what our government is up to *ought* to be available to the public. When this is information is available and widely interconnected the public has a deeper understanding of how our government works and gets more engaged. Millions more Americans are getting a deeper exposure to the workings of their government than they would be without this kind of information. Being able to read the bill that Congress is about to pass is a little like the experience of being in the capital for the first time and seeing the Declaration of Independence under glass protecting it. Our nation isn't abstract. There it is on paper right there.

This is all broadly data visualization. One of the most interesting recent developments in data visualization I've seen is a new prototype developed by Martynas Jusevicius called SemanticReports.com which is like a ManyEyes for the Semantic Web. It generates visualizations based on data published in the Semantic Web data format. Using Semantic Reports I was able to generate a map of campaign contributions to a congressman based on government data I had put into the Semantic Web years before.[3] (That was 18 million triples of campaign finance data and 1 billion triples of U.S. Census data.) This is a great way to explore just how much campaign cash comes from outside a legislator's district. But Semantic Reports isn't for the faint of heart. You have to enter a text-based query into Semantic Reports to generate the visualization, which means you need to be proficient both in cutting edge Semantic Web standards, in this case SPARQL, and in the ontology of the data being visualized, which again means you have to be proficient in the Semantic Web data model. But I love how it is entirely agnostic as to what the data is about *or how it is structured*. It's one of the few visualization tools that I've ever seen that doesn't require you to put your data into a table first.

I guess we were beaten to the punch by Sweden 250 years ago, but in the last five years government data geeks like me have been coming together to figure out our ask. I've been begging the Library of Congress to free their legislative data for three years now and haven't gotten a satisfactory response. In fact, in 2009 Congressman Honda of California took interest in the issue and had some legislative language passed that was supposed to push things along. Usually people joke that they need Act of Congress to get something done. Well I actually got an Act of Congress and still nothing has changed. If any of Newton's Laws applies to government, it is his law of inertia. Some day this has got to end.

Let's put aside what data we want. But, how do we want it? A group of techies came together in 2007 and formed the 8 Principles of Open Government Data, but a number of other groups have been thinking about this independently. Notably, the Open Knowledge Foundation in the UK has been on top of open data for a while. Unfortunately, their principles tend to be weaker than many would impose on a government, rather than say a private entity, so I've somewhat unfortunately sidelined them in these slides.

---

3 http://semanticreports.com/reports/d5e357a0-575d-42c4-9451-4ab9ea6ee86c

In a review of all of this work, I found some sixteen principles or best practices that have been proposed for what "open government data" means. Starting with the 8 Principles, public government data should be complete and primary, basically meaning not aggregated; it should be timely enough that it is still relevant to any ongoing policy debates; it should be accessible in both a disability and digital sense; and perhaps most importantly it should be machine-processable so that tools can be built to search, sort, and transform the data into other form. Records of congressional committee votes are regularly posted as scanned images, which all but prevent automated analysis of the information. Last year the House of Representatives published detailed internal spending data in the format of a several-hundred-pages long PDF. It wasn't scanned images, at least, but there's no way to know whether the methods developers came up with to extract the tabular information was reliable, not knowing how the PDF was generated. We're facing an uphill battle getting acceptance of just well-document CSV or XML, or even a flavor of PDF called Tagged PDF might be okay, but no one seems to know about it.

Government data should be available on a non-discriminatory and license-free basis and in a data format that is not encumbered by patents. This is where things start to diverge from the Open Knowledge Foundation's Open Knowledge Definition, or OKD. The OKD allows for a license. And when it comes to non-government works that sounds right. But in U.S. culture where the government can't copyright works and where any restriction on the free flow of information comes under high scrutiny, any license restricting the use of government published information is completely unacceptable --- except where the information might impinge on private or security or other narrowly defined exceptions. The European culture is different. In countries with a crown copyright, it might seem less onerous for a government to tell you not just whether you are permitted to share any arbitrary government-copyrighted data but also *how* you are permitted to do so according to the license's terms. I'll be a moral relativist on this one and let that slide.

Government data should also be online, free, and at a permanent address. The US ACM recommends it should be presented in a format that promotes analysis and reuse, does not impose computer security risks to users, and retains provenance through digital signatures. Others have suggested it be crafted with public input and inter-agency coordination and is subject to ongoing public review, uses globally unique identifiers for records, is web searchable, and is published with Linked Open Data methods.

The principles can serve as a benchmark. When the Obama Administration finished its Open Government Directive at the end of 2009 --- it was a memorandum that set goals for federal agencies --- we saw many of our data needs addressed. In fact to my surprise, the directive addressed nearly all of these open government data principles, and essentially added two of its own: being pro-active about data release and creating accountability by designating an official responsible for data quality.[4]

But this is rare, and it would be a mistake to take openness to be a binary value. Governments are storied institutions with long histories, deep existing infrastructure, and competing priorities and demands. It is impossible to expect that governments could make all data instantly open on every dimension. And that's fine.

So for instance, if you watch some streaming video from the U.S. House of Representatives, you are first greeted with a note not unlike the scary FBI warnings that used to start off all VHS tapes. "The use of duplications of broadcast coverage of the Committee on Transportation is governed by the rules of the House. Use for political or commercial purposes is expressly prohibited." Political speech is the most sacred of all speech in the United States. Could it be true that we are actually prohibited from using certain government records in political speech? I don't think this message was ever intended for

4 http://razor.occams.info/blog/2009/12/09/open-government-directive-evaluation-on-principles/

citizens, and the notice aside I doubt there is any legal prohibition here. Rather, my best guess is that it was originally intended to prohibit Congressmen themselves and the media who are registered in the press gallery from using the videos in unbecoming ways.

In some cases we're actually seeing substantive efforts by legislators to codify some of these principles as law. A bill before the New York State Assembly actually includes the text of the 8 Principles of Open Government Data. Not that I expect these principles to really end up codified as law. It would be nice, but like I said --- this isn't a binary thing. There are a lot of best practices for government data and we're just at the beginning of figuring out how to work with our governments on implementing each them.

But getting back to the beginning: what does open data buy us? During the debate over Kenya's constitution in 2005, political leaders faced the problem of informing a significantly illiterate electorate about the choices they faced. To help them at the polls, oranges became the symbol of a no vote while bananas the symbol for a yes vote.[5] It's more than a digital divide that prevents raw data from having some immediate value, and there are more immediate policy issues that the public needs to be engaged in than open data. By and large, the general public cannot do anything with raw data. Maybe some day when data tools evolve in the direction of Semantic Reports which I showed earlier. Until then, it's up to developers and statisticians and other data practitioners to help the public make sense of government data.

New data on its own can have unintended consequences. Michael Gurstein wrote in a blog post[6]: "Newly available access to land ownership and title information in Bangalore was primarily being put to use by middle and upper income people and by corporations to gain ownership of land from the marginalized and the poor." Gurstein points out that not all data yields an "effective use" of the data. Clay Shirky has claimed[7] that transparency fueled the lobbying business. I don't know if that's true, it's a little before my time, but I found LobbyData.com. From $50 to $500 per month you can get profiles of lobbying firms to help you find your next client or do competitor research. New levels of transparency can also affect the procedures within government in unexpected ways. The more we can see of Congress, the more congressmen want to take their negotiations somewhere else more private. And by no means is that necessarily a bad thing: some negotiations need that off-line personal touch. I call this the Wonderlich Transparency Paradox named after my colleague John Wonderlich at the Sunlight Foundation. He said: "How ever far back in the process you require public scrutiny, the real negotiations . . . will continue fervently to exactly that point."[8]

This is why I said transparency is a paradox at the start. We can't always use data to pry a closed government open. And data by itself doesn't necessarily empower citizens, and if it does it might empower the wrong ones. Data is not the end goal, it's only a way-point on the way to education, journalistic reporting, and computer applications.

But I will end saying that maybe transparency is *sometimes* an end to itself. If you saw the movie The Matrix, remember that Neo had to choose between leading a fake life, ignorant of how the world really was. Or he could leave the Matrix but face a more challenging and sometimes horrifying life. He chose to face the truth.

---

5 http://www.nytimes.com/2005/10/16/international/africa/16kenya.html?_r=1

6 http://gurstein.wordpress.com/2010/09/02/open-data-empowering-the-empowered-or-effective-data-use-for-everyone/

7 http://groups.google.com/group/openhouseproject/msg/53867cab80ed4be9

8 http://groups.google.com/group/openhouseproject/msg/94060a876083d86a