

CHAPTER EIGHTEEN

Case Study: GovTrack.us

Joshua Tauberer

More than 10,000 bills are on the table for discussion in the U.S. Congress at any given time. The most important bills can be 500 pages long or longer, and they can be rewritten several times on their way to becoming law. And with its 200 committees and subcommittees, Congress is overwhelming to anyone who is watching. The sheer volume of information coming out of Congress is itself a threat to government transparency. How can a representative be accountable if his legislative actions are too numerous to track? How can one take a stand on a bill if it is impossible to find? How can one know the law when it takes days to read a single bill?

Technology is a key player in government transparency. It's our own defense against the threat of government information overload. Looking for a bill? Do a search. Following a bill? Get a computer to track its changes.

Innovating the public's engagement with Congress has been the motivation behind GovTrack.us, the free Congress-tracking website that I built and have been running since 2004 (mostly in my spare time). GovTrack gathers the status of legislation, voting records, and other congressional information from official government websites and then applies the latest technology to make the information more accessible and powerful. Today the site reaches around 1 million people each month—visitors to GovTrack directly as well as visitors to other sites that reuse the open legislative database that GovTrack assembles. Figure 18-1 shows a screenshot of the GovTrack website.

Govtrack.us
a civic project to track Congress

Home Track Research Events Blog Tools About Dev

Bill Search: bill number or keywords

SHARE

Congress > Legislation

H.R. 2454: American Clean Energy and Security Act of 2009

111th Congress
2009-2010

To create clean energy jobs, achieve energy independence, reduce global warming pollution and transition to a clean energy economy.

Overview

Sponsor: Rep. Henry Waxman [D-CA30] [show cosponsors \(1\)](#)

Text: [Summary](#) | [Full Text](#)

Status:

- Introduced May 15, 2009
- Referred to Committee [View Committee Assignments](#)
- Reported by Committee May 21, 2009
- Amendments (2 proposed) [View Amendments](#)
- Passed House Jun 26, 2009
- Senate Vote -
- Signed by President -

This bill has been passed in the House. The bill now goes on to be voted on in the Senate. Keep in mind that debate may be taking place on a companion bill in the Senate, rather than on this particular bill. [Last Updated: Sep 18, 2009 8:42PM]

Last Action: Jul 7, 2009: Read the second time. Placed on Senate Legislative Calendar under General Orders. Calendar No. 97.

Related: See the [Related Legislation](#) page for other bills related to this one and a list of subject terms that have been applied to this bill. Sometimes the text of one bill or resolution is incorporated into another, and in those cases the original bill or resolution, as it would appear here, would seem to be abandoned.

Votes: Jun 26, 2009: This bill **passed in the House of Representatives** by roll call vote. The totals were 219 Ayes, 212 Nays, 3 Present/Not Voting. [Vote Details](#).

Votes of representatives you are tracking:
Rep. Chaka Fattah [D-PA2] voted Aye.

[View all 2 votes](#) on this bill.

Question & Answer

Can you answer any of these questions posed by other users? Think of it as a civic good deed. See 48 more questions posed on this topic or submit your own question on the [Q&A page](#).

Sep 29, 2009 12:56 AM - How will REDD projects be considered? Is it possible for REDD projects to deliver fully fungible credits into the ETS? - [Answer it!](#)

Oct 3, 2009 9:28 PM - We live in a moved on site home in the 1970 and are still on well and septic. Where does that leave us? - [Answer it!](#)

[and 48 more questions.](#)

Sources of Influence

[MAPLight.org](#) reports that the following organizations have taken a stance on this bill:

Support	Oppose
One Sky	American Farm Bureau Federation
Environment America	Greenpeace
League of Conservation Voters	Rainforest Action Network
VoteVets.org	Public Citizen
Natural Resources Defense Council	international rivers

Follow the link to [MAPLight.org](#) to see if campaign contributions from employees of these organizations are correlated with how Members of Congress voted on this bill.

Because the U.S. Congress posts most legislative information online one legislative day after events occur, GovTrack is usually one legislative day behind. For more information about where this data comes from, see [About GovTrack.us](#).

To cite this information, click a citation format for a suggestion: [APA](#) | [MLA](#) | [Wikipedia Template](#)

GovTrack.us is a project of Civic Impulse, LLC. Read about GovTrack. Feedback (but not political opinion) is welcome to operations@govtrack.us, but I can't do your research for you, nor can I pass on messages to Members of Congress. This site is "copy left". You are encouraged to reuse any material on this site. [Developers](#). GovTrack is open source and supports open knowledge.

Open Source

Navigation

- > Overview
- [Summary \(CBS\)](#)
- [Votes](#)
- [Full Text](#)
- [Committee Assignments](#)
- [Amendments](#)
- [Floor Speeches](#)
- [Reports](#)
- [Related Legislation](#)

Track H.R. 2454

This feed includes all major activity on this bill and its amendments, references in the Congressional Record, and relevant upcoming committee meetings.

[Preview Feed >](#)

Personalize your [Tracked Events](#) page and [email updates](#) by selecting trackers.

[Add Tracker](#)

Make a [widget for this tracker](#) to display on your web page.

Make a [widget that shows the status of this bill](#) for your webpage.

Primary Source

See [H.R. 2454 on THOMAS](#) for the official source of information on this bill or resolution.

FIGURE 18-1. The GovTrack website

Opening Legislative Data

Back in 1994, the newly elected Speaker of the House, Republican Newt Gingrich, pushed forward a revolutionary idea: use the Internet to keep the public as informed about the inner workings of Congress as the members of Congress were. And so the Library of Congress (LOC) immediately built THOMAS (<http://thomas.loc.gov>), where the public could, and still can today, find the status and text of all legislation in Congress (see Figure 18-2). As it still runs on 1990s technology THOMAS's innovation stops at keyword search. This is where GovTrack begins.

GovTrack integrates and cross-references more information than is on THOMAS, including voting records, biographical information on members of Congress, geographic information on congressional districts, and cost estimates from the Congressional Budget Office. By bringing everything into one place and cross-referencing the data, GovTrack is also able to compute novel statistics about the legislative history of each member of Congress, including an ideological score based on cosponsorship patterns and leader-follower scores based on who cosponsors whose bills. The site's presentation is tuned for a wide audience, and it allows users to personalize their view into Congress. It's also a community of wonks helping each other understand Congress. And the legislative database that powers the site is shared openly with other websites that mix and mash the information in new ways.

The LIBRARY of CONGRESS THOMAS

The Library of Congress > THOMAS Home > Bills, Resolutions > Search Results

[NEW SEARCH](#) | [HOME](#) | [HELP](#)

H.R.2454
Title: To create clean energy jobs, achieve energy independence, reduce global warming pollution and transition to a clean energy economy.
Sponsor: [Rep Waxman, Henry A.](#) [CA-30] (introduced 5/15/2009) [Cosponsors](#) (1)
Related Bills: [H.RES.587](#), [H.R.2998](#)
Latest Major Action: 7/7/2009 Read the second time. Placed on Senate Legislative Calendar under General Orders. Calendar No. 97.
House Reports: [111-137](#) Part 1

All Information (except text)	Text of Legislation	CRS Summary	Major Congressional Actions
Titles	Cosponsors (1)	Committees	All Congressional Actions
Related Bills	Amendments	Related Committee Documents	All Congressional Actions with Amendments
CBO Cost Estimates	Subjects		With links to <i>Congressional Record</i> pages, votes, reports

THOMAS Home | Contact | Accessibility | Legal | FirstGov FEEDBACK

FIGURE 18-2. THOMAS.gov, the Library of Congress's legislative tracking website

The site's RSS feeds keep visitors up-to-date on what is happening in Congress. There are literally tens of thousands of feeds on GovTrack, one for every bill, member of Congress, and topic area. Users can create personalized feeds by picking up "trackers" throughout the site:

the site then creates a customized feed based on the user's choices that is updated whenever legislative activities occur related to the trackers the user chose, such as a new bill being introduced or a vote taken. The tracked events can also be sent by email.

GovTrack applies technology that didn't exist in 1994 to legislative information:

- A congressional district map uses the Google Maps API to help visitors find their representatives by allowing them to zoom to street level (see Figure 18-3).
- Changes are tracked to bills as they move through the legislative process: text comparisons can be made between any two versions of a bill. (This is based loosely on the GNU diff program, a Unix command-line tool for comparing text files.)
- Permanent URLs trace back to particular paragraphs in bills, and there are a few embeddable widgets that can be used on other websites.
- Twitter hashtag recommendations for bills help us all tap into conversations happening elsewhere on the Web.

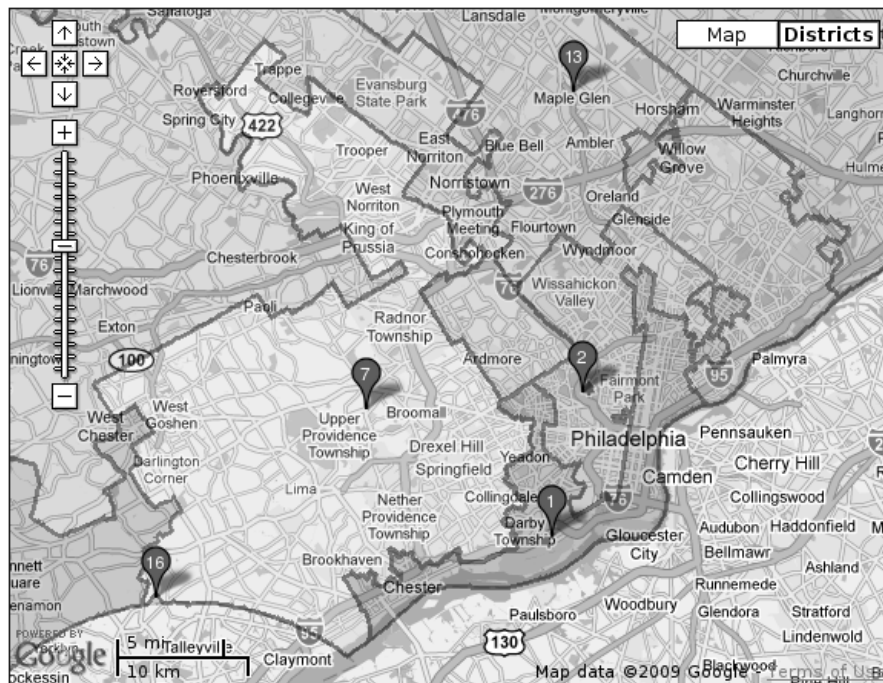


FIGURE 18-3. GovTrack's congressional district map is a mashup of Google Maps and cartographic data from the Census Bureau

Screen Scraping Congress

The best part is that the site runs itself. Almost all of the information GovTrack collects was already put online *somewhere* by the government. It's just scattered, difficult to understand, and harder yet to aggregate for the purposes of analysis. But this means that building a website such as GovTrack doesn't require continual manual labor to enter bill information into the system. Instead, I've programmed the site to periodically go out to government websites and fetch the information they have. It scans for new bill status daily, which is about as often as the THOMAS website is updated. The first goal, then, of GovTrack was to build a database of legislative information.

This was no small task, however, since most of the government websites weren't designed to share data with other websites. A human being might be able to make sense of a web page on THOMAS, but a computer doesn't know what's what on a web page. That didn't stop me, though: I just *told* the computer how to extract data from pages on THOMAS—how to look inside the HTML of the page, find keywords such as *sponsor*, then look to the right and pick out a name, and finally look up that name in my own database of congressional names to match it to the member of Congress it refers to (in past years, several have had the same name, which makes it trickier). This process is called *screen scraping*. My screen scrapers are a collection of small programs, or Perl scripts, that rely heavily on regular expressions and essentially manually written finite state machines to parse the structure of the pages.

It's not fun, and it's not perfect. Screen scraper scripts are not particularly difficult to write, but writing them does require a lot of trial and error. Because there is no guide to *everything* one can expect to see on THOMAS, I could only hope that I had considered all of the cases. I found out several years after finishing the bill status screen scraper that a bill can be sponsored by a person as well as by a committee. This was a surprising case that I hadn't anticipated. If a page on THOMAS displays something the screen scraper wasn't expecting, the screen scraper might bail with an error message, but it also might not notice and miss information on the page or, worse, misunderstand what is on the page, leading to inaccurate information being displayed on GovTrack. And when government websites change small details about how they format pages, the screen scrapers can get hung up and won't see the keywords they are looking for.

Fortunately, then, government websites change as often as...well, often they don't change.

This wouldn't be necessary if all of the legislative information was made available to the public as a database that a machine could process, such as a spreadsheet. That would eliminate the need for screen scraping and help make sites such as GovTrack more timely and reliable. And it would be cheap and easy for Congress to do because *it already has the databases*.

WHY DO THIS?

I wish I could say I had all this insight about innovating civic engagement when I first started building GovTrack in 2001. In the beginning, I thought the point of technology in civics was to help the public hold elected officials directly accountable; that if the public could get a better glimpse of what was happening in Congress they would be able to take that information to the polls. The site's first slogan was "Knowledge About Government Is Power" and it included an image of the all-seeing eye found on the back of a \$1 bill, which represented the public keeping an eye on Congress.

I realize now that it may have been a bit foolish to think legislative information would yield accountability. The information overload we face is not just a matter of size. A 700-page anything is a problem, yes, but with a bill we're facing 700 pages of precise legal text with references to parts of law you've never heard of on issues few people have enough expertise to truly understand. Then there are amendments, and procedural motions, and a whole other level of detail to consider. We elect representatives to read and draft bills so that we don't have to. Accountability is important, but a type of accountability where citizens weigh in on each vote is impossible.

It is also unnecessary. The type of corruption we actually find in Congress is not as obvious as members of Congress wittingly voting against their constituents while the constituents aren't looking. If money has an influence, it is well before the time of a vote. It is who gets elected, how they get on powerful committees, and what bills get introduced.

The goals of GovTrack have changed over time accordingly. The focus now is on addressing the great divide between what most people think goes on in Congress and how it really works. We are taught in school that members of Congress write laws. Yet the most important bills seem to go through the process so quickly that no one could possibly have had time to parse them. A congressman's chief of staff recently told me that there was "no" general process by which a bill becomes a law. We've all been taught the parliamentary flowchart, but is its relevance to what actually goes on so insignificant? We need to find out.

When I started GovTrack I felt downright righteous that government data should be free. That's "free" in the sense of the open source movement, meaning available to all and without restrictions on how it can be used or shared with others. The LOC had the data I wanted, but didn't share it. Commercial services collected the data I wanted (often by paying people to watch C-SPAN), but they sell access to it. Government data should be free! And if they weren't going to share it, I was going to build the best darn database of congressional information I could and make it available to all—undercutting commercial services and outdoing the government if I could.

Congressional Mashups

A number of other websites now reuse GovTrack's open legislative database. One of the most notable of these is MAPLight.org. MAPLight draws on congressional voting data from GovTrack and campaign contribution data from the Center for Responsive Politics (OpenSecrets.org) to identify correlations between money and votes in Congress. OpenCongress.org is a more social version of GovTrack based on GovTrack data. Almost half of the entries to Sunlight Foundation's 2009 Apps for America contest (<http://www.sunlightlabs.com/contests/appsforamerica/>) drew on data from GovTrack. The winner, Filibusted.us, uses data from GovTrack to highlight the senators who most often voted against *cloture*—that is, which senator most often derails progress in the Senate through a form of filibuster; see Figure 18-4. (Most of the websites reusing GovTrack data, including MAPLight and OpenCongress, have something or other to do with Sunlight Foundation, a new leader in developing and funding projects along these lines.)

NOTE

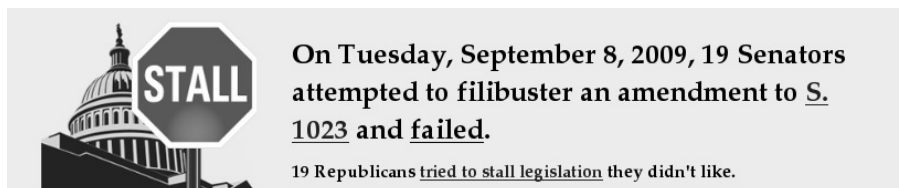
Other sites that reuse GovTrack's district maps widget include 50 Congress members' websites, the National Institutes of Health (NIH) recovery spending website, and My.BarackObama.com.

MAPLight.org is a particularly important example (see Chapter 20). Many uses of government data, including GovTrack, are attempts to provide something that some might say the government ought to be providing itself. But MAPLight.org is different because it does something we definitely wouldn't want the government to do: be its own watchdog.

There is no doubt that legislative data has uses and is of value to society in ways that we have not discovered yet. Each of these other sites contributes to getting the public more engaged in government and each educates the public in a unique way. The sites would be sorely missed if legislative data was not open: available online, in formats suitable for analysis and reuse, and free to be shared.

Changing Policy from the Outside

For the LOC to run its THOMAS website, it put together an XML database of legislative information. House and Senate clerks dutifully enter each day's proceedings into a computer, and this is what the public sees on THOMAS generally the next day. Not to diminish the great benefit to the public of THOMAS itself, but is that good enough? Given that the LOC is already putting together a legislative database, something that we know now can be the basis of many innovative and useful tools, what is their reason for not sharing it? If they shared that database in raw form with the public, building websites such as GovTrack would be much easier because screen scraping would no longer be necessary. Managers at the LOC have told me that they



On Tuesday, September 8, 2009, 19 Senators attempted to filibuster an amendment to S. 1023 and failed.

19 Republicans tried to stall legislation they didn't like.

Senate Roll Call #271:
 On the Cloture Motion (Upon Reconsideration Motion to Invoke Cloture on the Dorgan Amdt. No. 1347).
Vote passed, 80-19.

What does this mean?
 It takes **51** votes to pass a bill in the Senate. But it takes **60** votes to end debate. These days it's quite common for the minority party to threaten to *filibuster* (to debate indefinitely) in order to prevent action on a measure.
 A *cloture vote* is the Senate's way of asking "**Can we move on, please?**" It needs 60 votes to pass. There are 100 senators. As a result, as few as **41** senators can, as a bloc, bring the U.S. Senate to a standstill.
Still confused?

Why this bill?
 Good question. Read more about this bill and **decide for yourself** if it was worth holding up the business of the U.S. Senate.
 • Amendment: **Of a perfecting nature.**
 • Bill: **Travel Promotion Act of 2009**

Why Republicans?
 In the 111th Congress, Republicans tend to vote against cloture because they're in the minority. Historically, the party that is outnumbered wields the filibuster — but they've done so **more and more often** in recent years.

24
Votes

7.59%
Gridlocked

Scoreboard
 This was the **22nd** cloture vote of the 111th Congress.
 So far there have been **24** cloture votes out of **316** roll call votes — a percentage of **7.59%**.
 The 110th Congress (2007-2008) set a crazy record: **112** cloture votes out of **657** roll call votes. That's a percentage of **17.0%**! When it

Who voted against cloture?
 Is one of your senators in this list? Get in touch and ask what's up.
 AL [Jeff Sessions](#)
 AZ [Jon Kyl](#), [John McCain](#)
 ID [Mike Crapo](#), [Jim Risch](#)
 IA [Chuck Grassley](#)
 ME [Sam Brownback](#), [Pat Roberts](#)

FIGURE 18-4. *Filibusted.us*, created by Andrew Dupont for Sunlight Foundation's Apps for America contest, reuses GovTrack's legislative data to highlight the filibuster

have to balance sharing data against their responsibility to provide accurate information. And although this is a fair concern, it's not a difficult one to overcome.

Not only does the government not always share, as in the case of THOMAS, but sometimes they charge for data. The Government Printing Office (GPO) has had the audacity to sell some of its legislative and law data to the public at ridiculous prices (starting at \$8,000 per year per database), which is quite contrary to its mission, as given in law, to provide documents of value to the public at only the "marginal" cost of distributing them, meaning the cost of one additional copy. Today, the marginal cost of distributing most data on the Internet is \$0.00 (that's a tad less than \$8,000).

The GPO is an unusual case, however. The United States is generally a leader when it comes to not selling government information back to its citizens. The European model is cost-recovery plus *crown-copyright*. Citizens are charged a fee intended to let the government agency recoup costs (beyond the marginal cost), and the government asserts copyright over government publications to prevent citizens from redistributing them at a lower price. In the United States, distributing information at no more than the marginal cost is certainly the norm and

government works are generally not subject to copyright, which means that if you can afford to buy the documents from the GPO, you can give them away for free. Which I've tried to do.

What data the government should share and how it shares it is a policy question. However, policy can change. An active government transparency community, as well as interested congressional staff members (as part of Sunlight Foundation's Open House and Open Senate projects), are working toward changing policies that restrict open data. In September 2008, a number of us, including Carl Malamud of Public.Resource.Org, bought the Code of Federal Regulations (CFR; one of the components of U.S. law) from the GPO for \$17,325 with the intention to make it easier for other civic hackers to access the law of the land. Even though you can read the CFR on the GPO website, we wanted the underlying datafiles so that anyone can freely read and use the CFR as he likes. In 2009, under pressure from Malamud, congressional offices, and the White House's open government directive, the GPO finally began to acknowledge the value to the public of free access to raw materials. The CFR will likely be freely available by the time you read this.

Like the LOC, the Senate has an XML database of its roll call vote records. An XML database of roll call votes makes it easier to build visualizations of voting records like GovTrack and the *New York Times* do, for instance. I didn't know this for sure at the start, but it was a reasonable bet that any website with a lot of records stores its information in a database, and other pages on the Senate website about Senate committees had mistakes in the HTML (non-HTML XML tags showing through) that suggested that the information came from an XML database. (I had to look closely at the HTML to screen scrape it to form my own XML database.) Unlike the House, which has made roll call vote records available in XML since 2003, the Senate actually had a policy forbidding its webmaster from doing the same (something about senators reserving the right to inform their constituents as they saw fit). When I spoke with the director of law library services at the LOC about THOMAS and the Senate's webmaster about roll call votes, I didn't get the impression at all that either actually thought sharing their legislative databases with the public was a bad idea.

However, when I started GovTrack, neither the LOC nor the Senate webmaster had the appropriate mandate from Congress to share data, which was the biggest hurdle. In 2009, the Senate Rules Committee, which governs the Senate website, changed its policy on XML roll call votes (see Figure 18-5). This happened only after some prodding by myself, others inside and outside Congress, and finally, Senator Jim DeMint.

Another win for open data came in the Omnibus Appropriations Act of 2009. This bill directed the LOC to assess providing bulk, raw access to its legislative data. The provision was the result of The Open House Project Report (<http://www.theopenhouseproject.com>), which provided the House of Representatives guidelines on how to use technology better in the interests of transparency. Luckily, Rep. Mike Honda (D-Calif.) took an interest in this and had a bulk-data paragraph inserted into the legislative language. The LOC has not followed through yet, however.

```

- <roll_call_vote>
  <congress>111 </congress>
  <session>1 </session>
  <congress_year>2009 </congress_year>
  <vote_number>319 </vote_number>
  <vote_date>October 8, 2009, 04:26 PM </vote_date>
  <modify_date>October 8, 2009, 04:51 PM </modify_date>
  <vote_question_text>On the Motion to Recommit H.R. 2847 </vote_question_text>
- <vote_document_text>
  A bill making appropriations for the Departments of Commerce and Justice, and S
  the fiscal year ending September 30, 2010, and for other purposes.
  </vote_document_text>
  <vote_result_text>Motion to Recommit Rejected (33-65) </vote_result_text>
  <question>On the Motion to Recommit </question>
- <vote_title>
  Ensign Motion to Recommit to H.R. 2847 to the Committee of Appropriations
  </vote_title>
  <majority_requirement>1/2 </majority_requirement>
  <vote_result>Motion to Recommit Rejected </vote_result>
- <document>
  <document_congress>111 </document_congress>
  <document_type>H.R. </document_type>
  <document_number>2847 </document_number>
  <document_name>H.R. 2847 </document_name>
  <document_title>
  A bill making appropriations for the Departments of Commerce and Justice, and
  for the fiscal year ending September 30, 2010, and for other purposes.
  </document_title>

```

FIGURE 18-5. The Senate began publishing voting records in XML format in 2009, which helps sites such as GovTrack display the information in new ways

The what and how of open government data is increasingly becoming a mainstream policy question. Government transparency groups are suggesting open government data principles to guide policy (<http://razor.occams.info/pubdocs/opendataciviccapital.html>), but policymakers are taking the movement more seriously as well.

Engaging the GovTrack Community

The focus of GovTrack so far has been on providing comprehensive no-nonsense reference and tracking for legislative information. One of my long-term goals has been to build a community. Last year a GovTrack user asked me a procedural question about how Congress works: *can members of Congress change their vote?* (In the Senate, yes, with some restrictions.) It made me realize that there are many simple questions that people would like to ask about Congress or bills in Congress but have no one to turn to. Crowdsourcing could solve this problem. GovTrack users could gain from the wisdom of the crowd by having other users answer their questions. A user in the community might be an expert on a subject or be willing to do some legwork to find out the answers, such as reading the text of a bill.

And crowdsourcing worked. No sooner did I add question-and-answer boxes to pages for bills that visitors started asking questions and answering them. A visitor can post a question without logging in, as well as answer already posted questions. More than 7,500 substantive questions and answers have been posted over the past year (that's about one interaction for every 1,000 visits to the site). The Q&A submissions are moderated to ensure that users stay on topic. One of the most-answered questions has been "How will this bill impact day traders who may trade dozens to hundreds of trades per day?"

WHAT'S NEXT FOR GOVTRACK

I'm launching a Citizen Reporters experiment. The goal is to open the proceedings of congressional committees to the public. Committees are most visible when they hold hearings, which, when widely televised, are often a stage for posturing, but committee meetings are actually the locus of much of the legislating in the Capitol. Bills live or die at the mercy of committee chairs, and they also can be radically altered by committee member negotiations. The Citizen Reporters will go to public committee markup meetings and write up the proceedings of these meetings on GovTrack.

Other community tools exist on the sites reusing GovTrack's data. OpenCongress.org adds basic social networking, discussion forums, and a wiki. *Represented By* and *Laws I Like* are Facebook applications based on GovTrack data that let users track representatives and bills from within the social networking site.

Conclusion

The new guiding direction of GovTrack is to use technology to subtly change the dynamics of the system by helping the public to get a deeper understanding of how their government works. Congress will make better decisions when the public is more engaged, and the public can be more engaged if they better understand how Congress works. It's not about accountability so much as it is about education.

This is a learning process, though. The technologists have to figure out how Congress works before we can help others appreciate how complex it is. And we're just getting started.

About the Author



JOSHUA TAUBERER began GovTrack.us in his spare time in 2001 and is the director of Civic Impulse, LLC, a company he started in 2009. He is a software developer and is also finishing a Ph.D. in linguistics at the University of Pennsylvania. He holds a B.A. from Princeton University and an M.A. from the University of Pennsylvania.