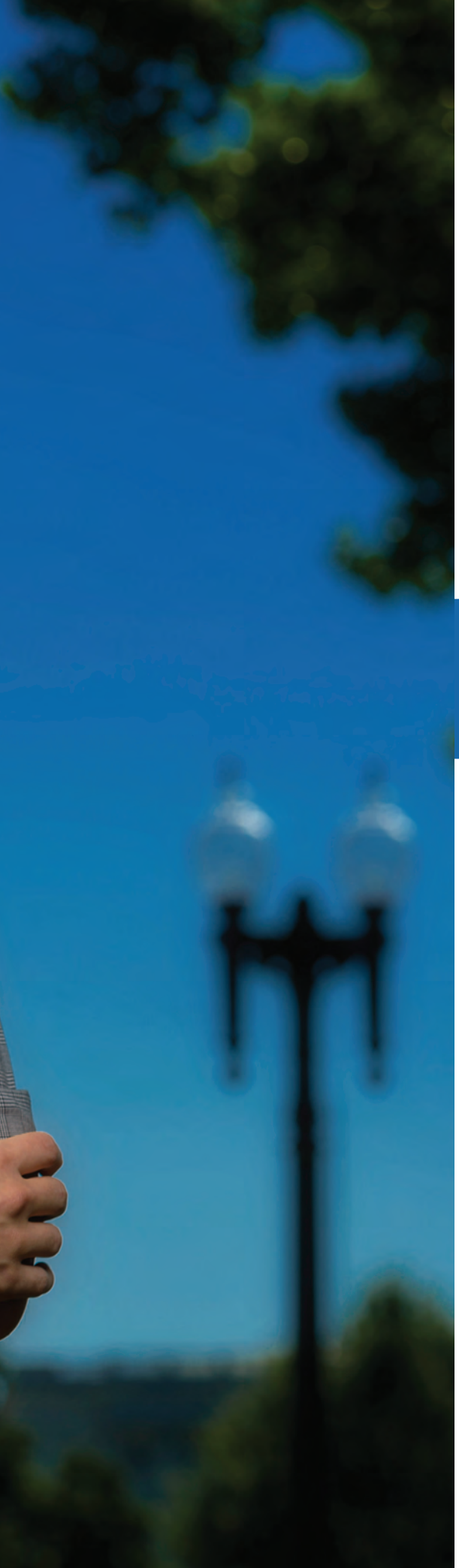

**Josh Tauberer Gr'11 believed the government should just give him the data he needed to create GovTrack, his website to help people follow the progress of Congressional legislation. When the powers-that-be said *No*, he went out and got it anyway.** BY ALYSON KRUEGER

# Civic Hacker

If you want to understand the 21st-century struggle to get the US government to release its data to the public, says Jim Harper, director of information policy studies at the libertarian Cato Institute, think about a newspaper.

Go to the weather page, and "you've got a lot of data," he says. "The maps, and the charts, and things like that, that people use all the time to assess what is up with the weather." Same for the financial section: "Data, data, data—really high numbers of facts per square inch, if you will." And don't even mention sports.

Then go to the national news. What do you see? Plenty of editorials and narratives rehashing what's "happening" on Capitol Hill, but very little hard information. "Data is really about representing facts," Harper says. "There isn't much to do with ideology, and everybody agrees that there should be more data so there could be less ideology, frankly."

One of the most prominent figures in this quest—Harper calls him "the guy" on open government data—is "civic hacker" Joshua Tauberer Gr'11. Tauberer is the author of *Open Government Data: The Book* (available for download at opengovdata.io), a signatory to the "8 Principles of Open Data" created by 30 leading open government data advocates in 2007, and a lobbyist to—and sometime consultant for—both the House and the White House on transparency issues. But he is best known for being the developer and maintainer of the website GovTrack.us, the go-to source for legislation-related information for journalists, lobbyists, and political activists of every persuasion, and even members of Congress and their staffs.

PHOTOGRAPHY BY JIM GRAHAM

"It's surprising that one guy with an interest in data was able to do as much as Josh has," Harper adds. Tauberer's success in collecting data scattered across multiple government websites and collating it into useful form (known as "screen scraping"), and his diligence in helping to pressure the government to make more of its data publicly available, has put him "at the leading edge of a massive change in the way government works," Harper says.

You wouldn't necessarily know it to look at him.

In appearance, Tauberer is a classic (and self-described) "geek": scrawny figure, boyish face, formidable computer bag—easily the biggest in the atrium of the National Portrait Gallery in Washington, where we meet. He went to Princeton undergrad, majoring in psychology, and his Penn PhD is in linguistics. His speech is punctuated by long, considering pauses, and he expresses himself softly in thoughtful, well-constructed sentences. At 31 he still comes off as shy and modest: a mention of a *New York Times* story about him is quickly followed by, "but I'm sure nobody read it …"

All the same, while he may not be a "torches and pitchforks" kind of guy—as fellow open-data advocate Daniel Schuman, policy director for Citizens for Responsibility and Ethics in Washington, puts it—Tauberer has been at the forefront of a decade-long initiative to get the United States government to release information about how it works and what it does.

Tauberer calls this "the application of Big Data to civics."

He was still in college when he decided he wanted to create a website that tracked every move Congress made—when it introduced a new bill, when that bill entered and left committee, how the bill was altered, when a vote occurred, and so on. But first he had to get the information from the government, a process that proved much harder than he expected. The problem is that, while the government does technically release that information to the public, it does so in a scrambled form that neither humans nor computers can easily put together to use.

To get the information he needed for his website Tauberer started hacking into government data. He didn't just keep what he found for himself but released it for use by any other programmer who wanted to build their own website or app.

In this context, "*Hack* doesn't mean anything related to cyber-terrorism," Tauberer insists. "Since I studied linguistics, I can say from an educated position that it is a different word from the word that is used for all the other stuff." (This is probably as close as he comes to using his Penn degree professionally, by the way.)

Along with a community of like-minded data-lovers, Tauberer also started lobbying the government to change the way it releases information about itself.

GovTrack currently receives 30,000 visitors a day. The website is so successful that it is almost always the first item that comes up during a Google search on legislation or congressional activities. Government bodies, including the House of Representatives and The White House, are working towards implementing Tauberer's suggestions on improving transparency. His work has also been featured in *The New York Times* and *Forbes* magazine, which in December 2011 named Tauberer one of the most influential people under the age of 30 in the law and policy arena.

"I could tell you a story, but I don't want to come across as really ridiculous," Tauberer begins, after a characteristic pause. "So, in third grade, you're supposed to bring in a book to read. And I brought in a programming manual. It's not narrative, right? Every page is just one particular function that you can program, and I'm just flipping through the pages and learning the functions."

Tauberer's father, a civil engineer, taught him to program when he was seven or eight years old. In high school at Plainview-Old Bethpage John F. Kennedy High School in Long Island—"which apparently is one of the best high schools in the country," he says—Tauberer won a website-design competition for a site called "Webcytology" that he created with his best friend. Another friend came up with the name, he says, a mashup of *website* and *cytology* (unicellular biology).

"It was an educational website that included a cool simulation of unicellular life," he explains. "You would design your own single-celled organism by choosing things like how many mitochondria to put in it, and the website would put your organism in a simulated colony with other users' organisms, and you would watch it replicate over days and weeks."

But even before he got to college in 2000, Tauberer knew that he didn't just want to be a programmer. "Which is weird," he says. "I should do what I'm good at, but I didn't want to."

A class discussion about the 1998 Digital Millennium Copyright Act in a course on copyright law, free speech, and technology sparked the idea of applying his talents to government transparency. The DMCA was intended to prevent people from circumventing copyright protection measures but ended up penalizing innocent people, he says—for instance, by blocking them from transferring a legally purchased book or video from one device to another. "In the class we were being told all the reasons why it was not a great law, and it seemed like an obviously bad law," Tauberer says. "And I thought, if the American public had better information about what was happening in Congress we might actually be able to prevent bad laws from happening, or at least hold people accountable."

In his free time Tauberer started searching for ways to stay informed about Congress. The Library of Congress's website, THOMAS, established in 1995, listed the status of bills, but it was notoriously unstable. ("The links on THOMAS break after 5 or 10 minutes," says Daniel Schuman.) The House and Senate websites were supposed to provide voting records, but those weren't always updated or complete. At the time (2001), there certainly wasn't any one place where he could get all the information he was looking for—congressional voting records, summaries of bills, notifications of when a bill was changed.

So Tauberer decided to build one.

The first component of GovTrack was the actual website that allows users to stalk Congress's every move. Among its key features are research tools that allow the user to search for a bill on a particular subject area and get email updates every time something happens in that arena. So if you are a doctor, for example, you can stay abreast of every law related to medicine. It is also the only site that displays edited bills in a marked-up form,

similar to Microsoft Word's track changes feature. Before GovTrack, the only way to compare current and past versions of bills was to read them side-by-side and look for changes.

The site also provides Prognosis, a statistical analysis tool that calculates each new bill's chances in Congress. This is a boon in particular to activists with limited resources in deciding where to concentrate their efforts, helping to level the playing field. "He's giving other people the opportunity to have access to high-quality information," says Schuman. "You don't have to just be a very wealthy corporation to know what's going on. You can be anyone."

As an example, take H.R. 2397, the Department of Defense Appropriations Act, 2014. The site offers summaries of the bill from the Library of Congress and House Republicans, explains its current status—as of July 24, it was passed by the House and sent on to the Senate for consideration—and predicts its chance of passage: 20 percent.

When he launched the site in 2004, "I had no conception that [it] could be a career in any way," Tauberer says. "It was a website. It was interesting. It had a couple of people visiting it. It was losing money, not generating money, which it does now."

His expectations for the site were so low that 2004 was when he decided to enter Penn's doctoral program in linguistics, a subject he had always found interesting—and in many ways similar to programming. "A lot of linguistics and a lot of programming is information management," he says. "What are the relationships between things and how do you express those relationships in a simple but [as] correct way as possible?"

But despite Tauberer's initial doubts, GovTrack has been hugely popular. In 2012, the site had five million users, including journalists, lobbyists for businesses and other causes, and members of Congress. "Everybody uses it!" exclaims Schuman. "The current system that Congress has available [to track legislation] isn't particularly good, and what Josh has built is much much, much better."

Perhaps the greatest service performed by GovTrack is the database of political information that Tauberer has assembled to create it. While the federal government has a huge database related to the affairs of Congress (all the bills in both chambers, all the action on those bills, who serves on what committee, etc.), it doesn't release that information directly to the public. Instead, information is spread out in bits and pieces across multiple websites—THOMAS, the House and Senate websites, the Government Printing Office's Federal Digital System (gpo.gov/fdsys), and others.

"He's figured out how to unscramble the eggs and make the eggs available to everyone to use for free."

For an individual looking for information, this system can kind of work—if you're willing to take the time and trouble to search the various websites and piece things together. Not so much if you're trying to build your own website or app harnessing that data.

Tauberer initially tried to get access to the original database, reasoning that, as an American, he deserved to have information about his own government in any form he found useful. But this reasoning proved unconvincing to the agencies responsible for that information.

"We periodically receive requests such as yours, so I know the answer is no, we are not able to provide anyone with direct access to our data," a Library of Congress staffer replied in May 2001 to Tauberer's request. The letter went on to note that all material on THOMAS was in the public domain, and "no permission is required to use it," and concluded, "Good luck with your project."

The question ever since, Tauberer says, has been "Where did this no-data-sharing policy come from? To be honest, I'm still not sure." But what it boils down to, he adds, is an attitude that "the public can't be trusted to have more information about government. It's perverse. And un-American."

The letter also suggested that Tauberer could use "robots" to gather data from the site. This is the same thing as "screen scraping," Tauberer says, which is ultimately what he did do, using software that recognizes different pieces of information and then compiles it into a central location.

Screen scraping, by the way, isn't easy, says Tauberer, likening the process to the story of Humpty Dumpty. "It takes years to be able to do it accurately," he explains. "You never really know which pieces fit where, exactly, until things happen. So, a veto is a really rare occurrence. Until a veto occurs, you can't really tell how it is going to appear on Congress's website, so you can't really predict how to program for it. And then, once it occurs you have to scramble and figure out, 'OK, now how do I add this to the database?'"

Tauberer not only used the database to create GovTrack, but also made it available to any programmer or developer for use in their projects. "Many, many dozens of people, many hundreds now, have taken this data and built something or at least tried to build something" with it, he says.

For example, MAPLight.org, which tracks the correlations between votes and campaign contributions, uses GovTrack's data, as does Filibusted.us, which records which Senators filibuster the most bills. And the House Democratic caucus uses GovTrack's data to run the internal web portal that keeps track of their legislative agenda—this after getting the same *No* answer Tauberer did from the Library of Congress to a request for the database, which he calls "a real example of just how locked down the data in Congress has been."

"It's not just that Josh has gone and figured out how to unscramble the egg," says Schuman. "He's figured out how to unscramble the eggs and make the eggs available to everyone to use for free."

"He illustrates pro-bono probably better than anybody else I've seen," echoes the Cato Institute's Jim Harper. "He's doing it because it's interesting to him, and it's going to help other people. What more do you need?"

"Once I started looking for data, I was insulted that the information wasn't

available for free anywhere," says Tauberer, explaining his motivation. "I guess I got it into my head that there was a moral obligation for the government to make core information available."

**While Tauberer was conceiving and developing GovTrack,** a community of like-minded advocates was beginning to form across the country. As founder of one of the oldest and most influential open-data websites, he was immediately thought of as a leading figure in that group.

"When you think about open data in this particular web-facing context, Josh has basically been doing it for a very long time," says David Robinson, founding associate director of Princeton's Center for Information Technology Policy, whose company, Robinson+Yu, consults on public policy related to the Internet. "He has far more experience than most of the people who work on these issues have."

In 2007, Tauberer was among the 30 experts—including Harvard law professor and copyright-reform advocate Lawrence Lessig W'83 ["Constitutionalist in Cyberspace," November 1998] and Tim O'Reilly, founder of the O'Reilly Media Group—who met in Sebastopol, California, to compose the eight principles of open government data, a document that has guided government bodies from the state of New Hampshire to the Obama administration on how they should release their data.

Around the same time, Tauberer also participated in the Open House Project, a group formed at the request of then soon-to-be Speaker of the House Nancy Pelosi to advise the House of Representatives on how to make its data more transparent and open. In addition to helping write the report, Tauberer—then also in the midst of his doctoral studies in Philadelphia—traveled to Washington to present the findings. "C-SPAN covered it, which was fun ... and exciting, of course!" Tauberer says, though no action resulted.

Tauberer and other advocates—such as Schuman, who previously worked at The Sunlight Foundation, a leading activist organization on transparency issues—have continued to lobby the House to adopt their advice. In particular, they want to see the release of the original government database that Tauberer has been trying to get since 2001.

"This is not an issue where we lose on the facts, right?" says Schuman. "People have a right to this information; it's not expensive to make it available in a way that it should be; they should just do it!"

One obstacle, according to Schuman, is that the institutions that actually control the data, the Library of Congress and the Government Printing Office, are afraid that releasing them in raw form will hurt their budgets: if they make all their data available and don't do anything with it, why would the government fund them? "It's a stupid fear, but that's their fear," he says, adding, "they aren't known for their dynamic leadership."

The current political environment—in which "Congress runs by crisis," says Schuman—makes it even more difficult to get the government to pay attention to something so technically complex, and which is not a subject of immediate, and loud, public concern.

Still, there are some promising signs. A new website called Congress.gov, designed to replace the outdated THOMAS platform and similar to GovTrack in the sense that it offers a step-by-step guide to what is happening in Congress, is out in a beta version. The House of Representatives also held a task force last year on open government data (to which Tauberer submitted ideas), and is expected to release its recommendations sometime before the end of the year. Tauberer and Schuman are hopeful that those recommendations will include finally making the Library of Congress data available to the public.

And there is also some optimistic news coming from the White House, which has had a mixed record in this area. On President Obama's first day in office he published a memorandum on open government, which "was quite forward thinking, and it said we're going to focus on transparency and public participation and collaboration," says Tauberer.

Though there is still a lot more to do, he adds, in many ways the administration has made great progress toward that goal. They produced a website, Data.gov, which is a catalog of datasets compiled by the Executive Branch, and urged other agencies to follow suit. Tauberer actually helped build one of these websites—the Department of Health and Human Services' Health.gov, which gives the public access to the vast amounts of health data it collects. This is especially useful for medical researchers, who can now use the government's data in their studies.

And in May the White House issued a memorandum defining open data that uses the definition Tauberer helped write back in 2007, mandates that government agencies should make their data open, and calls for collecting data in a way that facilitates making it open (such as not mixing classified and non-classified information, which delays the process and makes it harder to release data to the public).

While these are certainly steps in the right direction, the open government data community still has a few concerns. Some relate to specific clauses in the White House's recent memorandum—which, for example, leaves a lot of room for agencies to withdraw information because of speculative rather than demonstrated national security concerns. Other worries relate to incidents that have occurred during the Obama administration that aren't great for transparency in general (the Associated Press phone-record and NSA scandals, Obama's fierce prosecutions of whistleblowers, the lack of information on drones, etc.).

Outside of lobbying, Tauberer has also completed other projects to build the open data community. Besides his 2012 book, which details his "principles for a transparent government and an engaged public," he co-founded POPVOX, a site that helps constituents share their opinion with Congress easily. He also sponsors hackathons that bring together developers to invent new apps and websites that might help open government data.

But even if Tauberer just ran GovTrack, he would be making a huge difference, says Harper, because he is getting us one step closer to having lots of facts and hard information about politics. "It helps guide the way. It's a model; he's sort of doing research and development for the House Clerk's office or the Government Printing Office," he says. "If they see that it's been done, they know that it can be done, and they feel pressure to do it. Because it might be a little bit embarrassing to have one guy with a website doing a better job publishing information than your own huge organization."◆

**Alyson Krueger C'07 is a freelance writer in New York and a frequent *Gazette* contributor.**