Benchmarks for Measuring Success for Legislative Data Transparency

Daniel Schuman, Policy Counsel at the Sunlight Foundation
February 2, 2012

These remarks were delivered at the House Legislative Data and Transparency Conference.

Thank you to Matt Lira and Steve Dwyer for the introduction, and to the House of Representatives for holding such an important and timely conference. This kind of event has been a long time in coming.

I must acknowledge the excellent panels that have been happening all day. And I would be remiss if I didn't commend the Committee on House Administration for adopting "standards for the electronic posting of house and committee documents and data," which are already transforming the House in a very positive way.

Because I'm limited to 10 minutes, let me briefly commend three documents to all of you which lay out a transparency vision in greater breath and detail than is possible here. They are the Open House Project Report, the Ten Principles for Opening Up Government Data, and the report from the Congressional Facebook Hackathon.

I've been asked to speak about benchmarks for measuring success in making legislative data available online. I feel like a kid in a candy store, but I will try to restrain myself. When I speak about the House, please construe my remarks as applying to the Senate and the legislative support agencies as well.

WHAT IS TRANSPARENCY FOR?

In determining benchmarks, it's incumbent on us to assess, at least briefly: what good is online transparency anyway? Here's how I see transparency adding value to our political process. It provides relevant information to decision-makers at the time they need it. It levels the playing field between the special interests and everyone else so we all have an equal opportunity to find out what's going on. It lets the American people and their elected representatives have a solid basis for a conversation about priorities. It helps congress work more efficiently, by eliminating redundancies and identifying bottlenecks. It allows the agencies to better understand what they're supposed to do. It helps businesses make money by improving their ability to predict government actions. And most importantly, transparency is the cornerstone of a democracy.

This is all pretty ethereal, so I'll get to the point. To the maximum extent possible, **legislative information must be available online, in real time, and in machine readable formats**. With the exception of internal deliberations protected by the speech or debate clause, or national security and some personnel matters, the Congress's business is the people's business. So let me break down this formulation of online, in real time, and in machine readable formats into concrete benchmarks.

ONLINE PUBLICATION

Publishing information online is a major hurdle in of itself. A lot of information isn't online, but instead is only available if you know the right person, or go to the right room and ask for a hardcopy, and so on. Should you have to know someone on staff to get a copy of the chairman's mark on a bill before it's voted on? Do we really want to make people trudge down to the House's legislative resource center to print out documents at 10 cents a page? It certainly cannot make any sense to have to request a CRS report through your representative or pay 20 bucks online to buy a copy.

Almost as bad as the failure to publish online is secrecy through obscurity. If information is locked inside an image file and not susceptible to a search engine, or is in an entirely random location, or is hidden on page 400 of the congressional record, it's not really helpful to anyone.

In addition, old information can be just as important as newly created information. For example, there's a huge gap in the availability of committee reports. Along the same lines, while ignorance of the law is no defense for a crime, the actual enactment of the law, known as the Statutes are Large, is not available online for a nearly 80-year period.

Let me offer some concrete benchmarks by which we can judge improvements on this.

1. The House of Representatives should conduct an audit of all the different types of information it produces and releases, including whether it's online, and where it can be found.

2. To the extent the House (or legislative support agencies) has information that is already in electronic format -- from the documents in the Clerk's office to CRS reports to hearing transcripts -- that information should be put online in whatever format its currently in. It's also worth considering whether legislative data should include sometimes released items like Dear Colleagues and Whip notices. We can worry later about improving how this information is made available, but just to start, put them online.

Real-time Publication

Moving on, let's now talk about real-time publication. This is the kind of idea that makes a lot of people uncomfortable, but I'd suggest a common-sense starting point: think about the time frame and context in which a document is used. An amendment that's going to be voted on in 2 hours needs to be online just as soon as it's drafted. A bill that's going to be voted on in 2 legislative days needs to go up pretty quickly as well. You should know about a committee hearing a week in advance. Other items, like the House disbursement reports, can take a little longer.
Don't get me wrong. The goal should be real-time publication for everything. But the evaluation of what that means in the short term can be context dependent. But that context changes if the document is originally created in digital format -- in that circumstances, there shouldn't be any wait.

Here are some benchmarks:

1. All committee reports, amendments, and bills should be available online as they are introduced. The House should monitor the lag time between introduction and when they appear on THOMAS or the committee websites. I've done this, and it can be a while before some bills show up. Evaluate the extent of the problem, and work to reduce it.

2. All hearing notices should be available online 7 days prior to the hearing.

3. Many committees are skirting House rules about publishing video of hearings. House appropriators are particularly guilty of this. The House should review whether meetings are being held in rooms where video capability exists natively or could be added through use of the House's video service, and pester the committees if they're opting out of recording. When only one meeting in a particular committee is going on at a time, it should be streamed online so long as it is open to the public. It's time to review behavior and start slapping some wrists. Perhaps the House should create a mechanism for the public to report on non-webcast hearings.

Machine Readability

So let's move on to discuss machine-readable formats. This is what really allows the idea of House of Representatives as a platform for democracy to succeed.

The biggest wish of many staffers is to be able to dynamically see how an amendment would modify a bill, how that bill would change the law, (and eventually how an agency would promulgate a regulation, how the courts interpret that regulation, and back to congress again.) Along the same lines, people looking at a bill want to know if there are other, similar bills, in this congress or in previous ones, whether there are committee reports, CRS and GAO evaluations, and so on. If you cannot find a way to tie this information together, this dream becomes impossible.

Legislative data needs to be released as highly structured data. In other words, a machine needs to be able to look at the content and "know" what it is looking at. This would require the use of languages like XML, which allows this kind of value-added context. But to make it work, we also need a way to uniquely describe people and bills and amendments and so on -- cleverly enough embodied in commonly-accepted unique identifiers. There are already tons of these identifiers being used, but the House needs to consistently and widely employ them.

Sometimes, structured language is used when creating a document, or unique identifiers are used to describe data items in a document, but that document is stripped naked before it is released to the public. There are some circumstances where this makes sense, like hiding the different internal drafts of a bill. But most of the time, it serves no real purpose. The data that's removed could be very helpful to those on the outside. Leave it in.

Let me add that PDFs, especially PDFs that are image files, do not promote transparency. They make it difficult to impossible to extract data from documents. If you must use a PDF, make sure that the underlying data is available some other way as well.

That brings me to a point about how the data is made available. A lot of transparency advocates build scrapers to try to transform data that's published online and put it back into a useful structure. Josh Tauburer, for example, scrapes THOMAS to turn it into a database. It's like trying to unscramble an egg. Legislative data, such as that in THOMAS, should be made available online in bulk. Give folks the database all at once or in very large chunks, and let them figure out how to use it.

Here are my benchmarks:

1. All bills, amendments, and votes should be published online in XML, or some other structured format. Make scrapers unnecessary.

2. End the tyranny of only publishing in PDFs. House expenditure reports are a giant database -- publish them as a spreadsheet file, not a PDF. The Constitution Annotated is prepared in XML, don't publish it as a PDF.

3. Encourage the use of unique identifiers, whether they come from inside the House or elsewhere. The data needs to be interoperable.

CONCLUDING REMARKS

My time is running short, so I will only make two more comments about process.

First, today's conference, and the standards released by the House in December, are a good thing.

As a benchmark, we need to have another conference like this one within the next year as a way of assessing how well we have done, and we should continue with these conferences on a regular basis.

Second, we need to foster collaboration between those inside and outside government. In particular, technologists who are trying to use legislative data need to be able to get technology questions answered by the responsible internal stakeholder. And policy works can help provide direction so that the new services developed by the House meet the needs of the public. I suggest:

1. The creation of a standing committee, composed of internal and external stakeholders, that meets at least quarterly, if not monthly, to discuss these issues.

2. A listserv where people who are not in DC can engage in this discussion with people inside and outside of government.

I appreciate your time and the opportunity to speak. Thank you very much.