Congressional Data Coalition Testimony, FY 2017
Submitted by Daniel Schuman, Policy Director, Demand Progress
March 22, 2015

Dear Chairman Graves, Ranking Member Wasserman Schultz, and Representatives Amodei, Rigell, Jenkins, Palazzo, Farr, and McCollum:

Thank you for the opportunity to submit testimony on legislative branch funding priorities for fiscal year 2017. We focus on ways to strengthen the House of Representatives capacity to govern effectively and to engage with the American people through great transparency.

About us

The Congressional Data Coalition is a coalition of citizens, public interest groups, libraries, trade associations, and businesses that champion greater government transparency through improved public access to and long-term preservation of congressional information.[1]

Recognition of Ongoing House Activities

We commend the House of Representatives for its ongoing efforts to open up congressional information. We applaud the House of Representatives for publishing online and in a structured data format bill text, status, and summary information—and are pleased the Senate has joined the effort. We commend the ongoing work on the Amendment Impact Program and efforts to modernize how committee hearings are published. We look forward to the release of House Rules and House Statement of Disbursements in structured data formats.

We would also like to recognize the growing Member and Congressional staff public engagement around innovation, civic technology and public data issues. From the 18 Members and dozens of staff participating in last year's nationwide series of #Hack4Congress civic hacking events to the Second Congressional Hackathon co-sponsored by House Majority Leader Kevin McCarthy and House Democratic Whip Steny Hoyer, there is a growing level of enthusiastic support inside the institution for building a better Congress with better technology and data.[2] Moreover, the House Ethics Committee's recent approval of open source software and the launch of the Congressional Open Source Caucus means good things are in store for 2016.

This groundswell of support cuts across all ages, geographic areas and demographics, both inside and outside Congress. We are excited for the House's 2016 legislative data and transparency conference and appreciate the quarterly public meetings of the Bulk Data Task Force.

Summary of Requests

- Extend and Broaden the Bulk Data Task Force
- Release the Digitized Historical Congressional Record and Publish Future Editions in XML
- Publish all Congress.gov Information in Bulk and in a Structured Data Format

---

[1] For more information, visit http://congressionaldata.org/.

[2] Initiatives such as the Congressional App Challenge—which in 2015 saw a record 1,722 students build 493 apps in 187 Congressional districts, a full 43% of the House—demonstrate that the House's efforts to open up itself and its information to the public are not only bearing fruit, but are gaining traction far beyond the realm of good government geeks.

- Include All Public Laws in Congress.gov
- Publish Calendar of Committee Activities in Congress.gov
- Complete and Auditable Bill Text
- CRS Annual Reports and Indices of CRS Reports
- House and Committee Rules
- Publish Bioguide in XML with a Change Log
- Constitution Annotated
- House Office and Support Agency Reports

Extend and Broaden the Bulk Data Task Force

One major success of the House's legislative modernization efforts was the creation of the Bulk Data Task Force,[3] the recommendations of which led to the online publication of bill summaries, status, and text in a structured data format, as well as other improvements. While the Task Force issued its final report in the 113th Congress, many of its participants continue to meet. The Task Force is a unique forum for congressional content creators and publishers to work together and interact with the public. We wish for it to be placed on a permanent footing before the end of the 114th Congress.

We urge the committee to formally reestablish the Task Force on a permanent basis and expand its mission to broadening availability of congressional information in machine readable formats. There is precedent for this, with the XML Working Group that was created in the 1990s to establish document type definitions for use in creating legislative documents in XML.[4] Its scope should include legislative information and records held by committees, offices, and legislative branch agencies as well as other information concerning the operation of Congress.

Release the Digitized Historical Congressional Record and Publish Future Editions in XML

The Congressional Record, as the official record of the proceedings and debates of the Congress, is central to understanding congressional activities. Many of the resources we have come to rely upon, such as Congress.gov, republish just a fraction of its contents. Unfortunately, the Congressional Record is not published in bulk in a structured data format, but instead as plain text, and, in some cases, as (even less versatile) PDFs. In addition, the Congressional Record is available online only from 1994 forward and prior to 1873.

We understand that the GPO has completed an effort to digitize issues of the Congressional Record to fill the 100-plus-year gap,[5] but will not release the digitized documents (PDFs) until metadata is added. This could be a lengthy process. We request the digitized records be released now; we can build our own metadata[6] and use the documents while the official metadata process finishes.

Going forward, we urge the Congressional Record be published in a structured data format, like XML. There have been efforts by the public to scrape the version of the Congressional Record,[7] but the results were incomplete, the same scrape-able information no longer exists on Congress.gov, and PDFs are less useful than structured data. We are sensitive to the cost constraints on GPO but suggest

---

[3] House Report 112-511, available at http://www.gpo.gov/fdsys/pkg/CRPT-112hrpt511/pdf/CRPT-112hrpt511.pdf.
[4] See http://xml.house.gov/
[5] See http://www.scribd.com/doc/48672433/Constitution-Annotated-Congressional-Record-and-Statutes-at-Large.
[6] For example, we did just that with the Statutes at Large, where it is possible to look up every public law enacted by Congress from 1789 to present. See http://legislink.org/us.
[7] https://sunlightfoundation.com/blog/2014/02/20/sample-the-new-a-la-carte-congressional-record-parser/

that publication in a more versatile format may lead to reduced print demands, improved internal efficiencies, and greater reuse and transformation of the Congressional Record into useful products.[8]

Publish all Congress.gov Information in Bulk and in a Structured Data Format
We recognize and appreciate all the hard work that went into publishing bill summaries, status information, and text in a structured data format from the 113th Congress forward. We urge the process be extended to all bill summaries, status information, and text held on Congress.gov.

To accomplish this and the other ends discussed elsewhere, we urge continued financial support for GPO and LC to maintain and develop congress.gov, fdsys.gov, and their successors.

Include All Public Laws in Congress.gov
It is vitally important for everyone to know what the law is. In many instances, the U.S. Code is not the law and yet the legislation signed by the president is not available to the public in an easy to find way. We respectfully suggest that every public law be available for retrieval through Congress.gov.

Publish Calendar of Committee Activities in Congress.gov
Thanks to the creation of docs.house.gov and a Senate information repository, it is now possible to systematically gather information on all upcoming committee hearings, markups, and other activities. This information should be combined into a single calendar on Congress.gov, with descriptions of upcoming activities, links to live and archived video, and links to relevant committee information.

Complete and Auditable Bill Text
The Government Publishing Office is charged to accurately and authentically print the bills before Congress, yet there are gaps in GPO's archive—as seen on FDSys—without any explanation. In addition, public access to the text of bills in the 101st and 102nd Congresses are being removed as a part of the retirement of THOMAS.gov. Furthermore, GPO holds structured data for bills prior to the 111th Congress (when both House and Senate legislation were first published in XML), which it does not make available to the public at all (locator code format). We ask that GPO publicly report on the presence or absence of public access to all prints of bills starting with the 101st Congress, including access to the prints in a structured data format, with a public audit log in CSV format. This would build trust in GPO's authenticity and accuracy processes.

CRS Annual Reports and Indices of CRS Reports
CRS reports often inform public debate. Its analyses are routinely cited in news reports, by the courts, in congressional debate, and by government watchdogs. We do not wish to weigh too deeply into the debate over public access to the reports in this forum[9] as there is legislation to effectuate a change in how they are made available to the public.[10] However, we wish to bring three issues to your attention.

---

[8] In the meanwhile, publication of the Congressional Record in locator code format along with GPO's locator code-to-PDF conversion software, in source code form, may suffice in the interim.
[9] There have been several letters of late urging public access to CRS reports from former CRS employees, a large coalition of civic organizations, and a coalition of conservative organizations. Many have endorsed the legislation as well. For a complete list of correspondence, see https://github.com/DanielSchuman/Policy/wiki/Congressional-Research-Service#members-of-congress-and-organizational-letters.
[10] See H.R. 4702, the Equal Access to Congressional Research Service Reports Act of 2016, available at https://www.congress.gov/bill/114th-congress/house-bill/4702.

First, CRS apparently is producing two versions of its annual report: one for publication online on CRS's website and another apparently for congressional staff.[11] The version released to the public makes no indication that it has been redacted or truncated. It appears the difference between the two reports is the non-truncated version lists CRS reports released during the prior year. We suggest that the American people be able to see the full report provided to Congress. These silent redactions are misleading and unnecessary. In addition, we suggest that CRS cease removing prior annual reports from its website.

Second, we request the Committee require contemporaneous online reporting of the names, report numbers, and publication/revision/withdrawal dates for CRS reports. We do not include CRS memoranda, which are confidential. In this way, members of the public may contact their representative if they see a report they are interested in upon its publication or revision.

Finally, we request the amending the language inserted into the legislative branch appropriations bill concerning expending funds for the purpose of publication so that it refers specifically to print publications.[12] The limiting language was put in place over concerns regarding printing and mailing costs and was initially inserted in 1954. Electronic publication of Congressional Research Service documents imposes no additional printing or mailing costs. It likely discourages CRS from contemplating whether any of its documents should be published online, which is a default position that no longer makes sense from a cost perspective.

House and Committee Rules

Crucial to understanding the House and its committees are their rules, but these vital documents are usually published as PDFs or garbled text files. The House rules for the 114th Congress are published by the Rules Committee but only as a PDF, and, if you can find it on FDSYS, it is available as a PDF file and an annotated, discontinuous TXT file. As another example, while the Committee on Rules makes its rules available as HTML, the Permanent Select Committee on Intelligence publishes its rules only as a PDF. Ideally, all rules should be published in a structured data format like XML. We understand that the House Rules are expected to be published as XML in the near future, but are unaware of plans for committee documents. Regardless, in the interim, in addition to however else they are published, rules should be published in an open, non-proprietary format, even if as a TXT, ODT or DOCX file, without the annotations that make GPO's version unusable for many purposes.

Publish Bioguide in XML with a Change Log

The Biographical Directory of the United States Congress (or Bioguide) is an excellent source of information about current and former members of Congress. Since 1998, the online version of the Bioguide has been maintained by staff in the Office of the Clerk's Office of History and Preservation and the Office of the Historian of the United States Senate at http://bioguide.congress.gov. And, since at least 2007, the underlying data structures for Bioguide data have been provided by the House at its

---

[11] See, for example, these two versions of the 2012 Annual Report. This version, released to the public, has 41 pages https://www.loc.gov/crsinfo/about/crs12_annrpt.pdf; and this version, available to congressional staff, has 101 pages http://www.scribd.com/doc/296990681/Annual-Report-of-the-Congressional-Research-Service-2012#scribd. The difference is the latter contains a list of CRS reports released in the instant year.

[12] "Provided, That no part of this appropriation may be used to pay any salary or expense in connection with any publication, or preparation of material therefor (except the Digest of Public General Bills), to be issued by the Library of Congress unless such publication has obtained prior approval of either the Committee on House Administration or the Senate Committee on Rules and Administration."

XML website. Unfortunately for those who wish to programmatically make use of the information, the website's data is published only in HTML. In addition, the Bioguide website provides up to three HTML files for each Member: a biography, extended bibliography, and research collection, which can triple the amount of work required to fully scrape the website. We recommend Bioguide information be published in XML. In addition, a change log for the Bioguide website through Twitter or an RSS/Atom feed would be helpful to keep the public apprised of updates/changes.

Constitution Annotated

The Constitution Annotated (or CONAN) is a continuously-updated century-old legal treatise that explains the Constitution as it has been interpreted by Supreme Court. While the Joint Committee on Printing required in November 2010 that GPO and CRS to publish CONAN online, with new features, and with updates as soon as they are prepared, it did not require publication in a machine-readable format.[13] This is an important omission, as the document is prepared in XML yet published online as a PDF, even while it is internally available to Congress as a series of HTML pages. In light of the House's broadening of documents availability in machine-readable formats, this issue is ripe for resolution. Publication of the XML source or the HTML pages would address some of these concerns.

House Office and Support Agency Reports

The legislative offices and agencies that support of the work of the House of Representatives issue annual or semi-annual reports on their work. These reports are of interest to the public, as they help explain legislative operations and often can help ensure public accountability. While some offices, such as the Chief Administrative Office, routinely publish their reports online, others do not, or do not do so in a timely fashion. We urge that the Committee to require all legislative support offices and agencies that regularly issue reports that summarize their activities to publish those reports online in a timely fashion, including back issues.

Thank you your attention to these issues. To discuss this further, please contact Daniel Schuman, co-chair, Congressional Data Coalition, at 202-577-6100 or daniel.schuman@gmail.com or Zach Graves, digital director, R Street Institute, at (202) 525-5717 or zgraves@rstreet.com.

Sincerely yours,

Center for Data Innovation
Data Coalition
Demand Progress
Free Government Information
GovTrack.Us
New America's Open Technology Institute
OpenGov Foundation
OpenTheGovernment.Org
R Street Institute
Sunlight Foundation

[13] See http://www.scribd.com/doc/48672433/Constitution-Annotated-Congressional-Record-and-Statutes-at-Large.

Daniel Schuman | Policy Director | Demand Progress

Testimony for the Record before the House Legislative Branch Appropriations Subcommittee

March 22, 2016

Daniel Schuman Biography

Daniel Schuman has long worked at the intersection of law, policy, and technology. He currently serves as policy director with ***Demand Progress***, a civic organization with 2 million affiliated activists that fights for fights for basic rights and freedoms needed for a modern democracy**.** He most recently served as policy director for ***Citizens for Responsibility and Ethics in Washington***, and prior to that as policy counsel with ***the Sunlight Foundation***. In a prior life he was a legislative attorney with the ***Congressional Research Service***.

Daniel chairs the ***Congressional Data Coalition***'s steering committee and founded the ***Advisory Committee on Transparency,*** which promoted the work of the Congressional Transparency Caucus. In 2013, Daniel was named among the "top 25 most influential people under 40 in gov and tech" by *FedScoop*. He is a nationally recognized expert on federal transparency and has testified before Congress and appeared on NPR, C-SPAN. Daniel graduated cum laude from Emory University School of Law.

[Insert Subcommittee Name Here]

Witness Disclosure Form

Clause 2(g) of rule XI of the Rules of the House of Representatives requires non-governmental witnesses to disclose to the Committee the following information. A non-governmental witness is any witness appearing on behalf of himself/herself or on behalf of an organization other than a federal agency, or a state, local or tribal government.

Your Name, Business Address, and Telephone Number:

Daniel Schuman, 30 Ritchie Ave., Silver Spring, MD, 202-577-6100.

1. Are you appearing on behalf of yourself or a non-governmental organization? Please list organization(s) you are representing.

Center for Data Innovation, Data Coalition, Demand Progress, Free Government Information, GovTrack.Us, New America's Open Technology Institute, OpenGov Foundation, OpenTheGovernment.org, R Street Institute, Sunlight Founndation

2. Have you or any organization you are representing received any Federal grants or contracts (including any subgrants or subcontracts) since October 1, 2012 related to the agencies or programs funded by the Subcommittee?

Yes [No]

3. Have you or any organization you are representing received any contracts or payments originating with a foreign government since October 1, 2012 related to the agencies or programs funded by the Subcommittee?

Yes [No]

4. If your response to question #2 and/or #3 is "Yes", please list the amount and source (by agency and program) of each Federal grant (or subgrant thereof) or contract (or subcontract thereof), and/or the amount and country of origin of any payment or contract originating with a foreign government. Please also indicate whether the recipient was you or the organization(s) you are representing.

Signature: Daniel Schuman Date: March 21, 2016