Congressional Data Coalition Testimony, FY 2016
Submitted by Daniel Schuman, Demand Progress Policy Director
Prepared for the Committee on Appropriations Legislative Branch Subcommittee
Regarding Appropriations for the Secretary of the Senate, the Library of Congress, the Government Publishing Office, and Sergeant at Arms

March 27, 2015

Dear Chairman Capito, Ranking Member Schatz, and Senators Kirk, Moran, and Murphy:

Thank you for the opportunity to submit testimony on legislative branch funding priorities for fiscal year 2016. Our recommendations focus on improving efficiency within and transparency concerning offices and agencies of the legislative branch, with an emphasis on better use of information technology.

About Us
The Congressional Data Coalition is a coalition of citizens, public interest groups, 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 Senate Activities
We commend the United States Senate for its recent commitment to publish bill status and summary information—soon to be joined by legislative text—online and in a structured data format. We also appreciate the quarterly public meetings hosted by the invaluable Bulk Data Task Force, of which delegates from the Senate often participate. We remain hopeful that progress will be made on the Joint Committee on Printing's directive to digitize volumes of the Congressional Record from 1873 to 1998.

Summary of Requests
- Extend and broaden the Bulk Data Task Force
- Publish the Congressional Record in XML and eliminate electronic publication gaps
- Publish a complete and auditable archive of bill text, in a structured electronic format
- Instantiate a Senate-wide committee record publishing system
- Publish a contemporaneous list of widely-distributed CRS reports that contains the report name, publication/revision/withdrawal date, and report ID number
- Release widely-distributed CRS reports to the public
- Publish Bioguide in XML with a change log
- Publish the Constitution Annotated in a machine-readable format
- Publish Senate office and support agency reports online
- Publish Senate Expenditure Reports in a machine-readable format

Extend and Broaden the Bulk Data Task Force
One of the greatest successes of the efforts to modernize legislative information was the creation of the Bulk Data Task Force,[2] the recommendations of which led to the online publication of bill summaries and text in a structured data format and the commitment to add bill status information this year, 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 hope the Senate will deepen its participation as it continues to send delegates from its Senate and legislative support offices to participate in deliberations.

---

[1] For more information, visit http://congressionaldata.org/.
[2] House Report 112-511, available at http://www.gpo.gov/fdsys/pkg/CRPT-112hrpt511/pdf/CRPT-112hrpt511.pdf.

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.[3] 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.

Congressional Record 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 less versatile PDFs. In addition, the Congressional Record is available online only from 1994 forward and prior to 1873. The Joint Committee on Printing authorized GPO to fill in the 100-plus-year gap in 2011,[4] although it is unclear whether online publication would be as structured data or in a less flexible format (such as PDF).

While there had been efforts by the public to scrape the version of the Congressional Record on the old THOMAS.gov,[5] the results were incomplete and the same scrapable information no longer exists on Congress.gov. Moreover, there is no substitute for official publication in a structured data format like XML. We urge the committee to inquire into GPO's efforts to fill the online publication gap and to require future publication of the Congressional Record in XML. [6]

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) that it does not make available to the public at all (i.e., in 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.

Instantiate a Senate-wide committee record publishing system

Committee documents are vital records of congressional activity, but they often are hard to find or search, and are subject to removal from a committee website when leadership turns over or websites are updated. We urge the Senate to institute a chamber-wide committee publishing system that serves as a comprehensive repository across committees and congresses.

To address this problem, the House of Representatives created Docs.house.gov, which "provides access to committee documents and text of legislation being considered in committee…" dating back to the 112th Congress in XML formats where available. It includes meeting notices, witness lists, witness and member statements, legislative and amendment text, and more. The Clerk of the House administers the site to ensure it is viewed as nonpartisan. Docs.house.gov guarantees that public access to committee records is

---

[3] See http://xml.house.gov/

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

[5] https://sunlightfoundation.com/blog/2014/02/20/sample-the-new-a-la-carte-congressional-record-parser/

[6] 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.

maintained even as leadership changes and committee websites are updated. We urge the Senate to provide the same level of access to its committee documents.

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. However, unlike its sister legislative branch agencies, CRS reports are not released to the public by CRS even though CRS routinely shares them with the media upon request and with officials in the executive branches. In addition, public access often is through third parties that routinely charge a fee for access. We believe all Americans should have an equal opportunity to be educated about important legislative issues—including knowing which reports have recently been released and having free access to them.

We request the Committee require CRS to contemporaneously publish online a list 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 senators if they see a report they are interested in upon its publication or revision. CRS already provides an annual report to the Committee, published on CRS's website, which lists the total number of reports issued or updated. In FY 2012, for example, 534 new reports were prepared and 2,702 reports were updated.[7] This accounting should be expanded to include an index of the reports and be updated on a daily basis in a machine-readable format.

We further request the public be provided direct online access to the recent Congressional Research Service reports, which we have discussed in prior testimony to the Committee.[8]

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. Since at least 2007, the underlying data structures for Bioguide data have been provided by the House at its 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.[9] 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. This issue is ripe for

---

[7] Annual Report of the Congressional Research Service of the Library of Congress for Fiscal Year 2012, p. 2, available at http://www.loc.gov/crsinfo/about/crs12_annrpt.pdf.

[8] See Comments of the Sunlight Foundation, May 24, 2013, available at https://s3.amazonaws.com/assets.sunlightfoundation.com/policy/testimony/Sunlight%20Foundation%20Leg%20Branch%20Approps%20Testimony%202013-05-24.pdf

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

resolution. At a minimum, publication of either the XML source or the HTML pages would address many of our concerns.

Senate Office and Support Agency Reports
The legislative offices and agencies that support the work of the United States Senate 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 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.

Semi-Annual Senate Report on Receipts and
The semi-annual Senate report on Receipts and Expenditures contain all spending by the U.S. Senate and are currently published online as a PDF. They should be published as data files, such as CSV, to allow for the public to easily analyze the information. The online publication that started in 2011 was a significant step forward, but the data should be available in a more flexible format.

We appreciate your attention to these issues. To discuss this further, please contact Daniel Schuman, policy director, Demand Progress, at 202-577-6100 or dschuman@citizensforethics.org or Zach Graves, digital director, R Street Institute, at (202) 525-5717 or zgraves@rstreet.org.

Sincerely yours,

Congressional Data Coalition
Data Transparency Coalition
Demand Progress
Free Government Information
GovTrack.us
OpenTheGovernment.org
R Street Institute
Sunlight Foundation