March 6, 2015

Rep. Tom Graves, Chairman
Appropriations Committee –
Subcommittee on the Legislative Branch
U.S. House of Representatives
Washington, DC 20515

Rep. Debbie Wasserman Schultz, Ranking Member
Appropriations Committee –
Subcommittee on the Legislative Branch
U.S. House of Representatives
Washington, DC 20515

Dear Chairman Graves and Ranking Member Wasserman Schultz,

Thank you for the opportunity to submit testimony on legislative branch funding priorities for fiscal year 2016. We focus on ways to further the House of Representatives' efforts to increase legislative transparency in accordance with the House of Representative's 114th Congress rules package's order on "Broadening Availability of Legislative Documents in Machine Readable Formats."[1]

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.[2]

Recognition of Ongoing House Activities
To begin, 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 status and summary information—soon to be joined by legislative text—and are pleased the Senate will join these efforts in the 114th Congress. In addition, the website http://docs.house.gov/ continues to serve as an excellent online source for committee and House floor information, thanks in large part to work performed by the Clerk of the House. Furthermore, the Rules Committee's website is a tremendous resource for learning about legislation to be considered on the House floor.

We also congratulate the Office of Law Revision Counsel for its ongoing improvements to publication of the US Code, which serve as a showcase of the potential of the House's efforts. We appreciate the House's annual conferences on legislative transparency and are looking forward to the 2015 conference. And we eagerly await the public roll-out of the Amendment Impact Program[3] and the LRC's codification tools as well as the quarterly public meetings

[1] H. Res 5, 114th Congress, Section C, Separate Orders, item (n), which states: "The Committee on House Administration, the Clerk, and other officers and officials of the House shall continue efforts to broaden the availability of legislative documents in machine readable formats in the One Hundred Fourteenth Congress in furtherance of the institutional priority of providing public availability and use of legislative information produced by the House and its committees."
[2] For more information, visit http://congressionaldata.org/.
[3] With AIP's automation of the consolidation of amendments into bills and bills into laws, we hope the public with be provided access to this and all of its software, in whole or part, through an application programming interface, to encourage third party developers to leverage this groundbreaking work and make legislation easier for the public to understand.

hosted by the invaluable Bulk Data Task Force. We also remain hopeful that progress will be made on the Joint Committee on Printing's obligation 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
- 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 the House rules and committee rules in a machine-readable format
- Publish Bioguide in XML with a change log
- Publish the Constitution Annotated in a machine-readable format
- Publish House office and support agency reports online
- Publish House Expenditure Reports in a machine-readable format

Extend and Broaden the Bulk Data Task Force

One of the greatest successes of the House's legislative modernization efforts was the creation of the Bulk Data Task Force,[4] 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 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.[5] 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 (even 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,[6] 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,[7] the results were incomplete, the same scrapable information no longer

[4] House Report 112-511, available at http://www.gpo.gov/fdsys/pkg/CRPT-112hrpt511/pdf/CRPT-112hrpt511.pdf.
[5] See http://xml.house.gov/
[6] See http://www.scribd.com/doc/48672433/Constitution-Annotated-Congressional-Record-and-Statutes-at-Large.
[7] https://sunlightfoundation.com/blog/2014/02/20/sample-the-new-a-la-carte-congressional-record-parser/

exists on Congress.gov, and 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. We are sensitive to the cost constraints on GPO but suggest 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]

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 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, and the most recent version of a report is not always available. We believe all Americans should have an equal opportunity to be educated about important legislative issues, and that includes 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 representative 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.[9] 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.

---

[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] 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.

In recent years CRS has declined to release its reports directly to the public in part based upon language inserted into the legislative branch appropriations bill.[10] That limiting language, however, was put in place over concerns regarding printing and mailing costs. Moreover, the modern language was initially inserted in 1954, 16 years prior to CRS' creation. A broad 1952 limitation on the Library of Congress was put in place because of concerns around printing costs.[11] In 1954 the language was loosed to allow publication with prior authorization by the Committee on House Administration or the Senate Rules Committee, but retained in part out of concerns of the cost of mailing the documents to "newspapers and women's clubs"[12] unless there was reimbursement for the costs of mailing.

Electronic publication of CRS reports imposes no additional printing or mailing costs. CRS already maintains a Congress-only website with reports published in an electronic format. Depending on how the reports would be released to the public—via FDSYS, via FTP, through a website maintained by the Clerk, through a GPO bulk data download,[13] or a website maintained by CRS—the costs would be minimal and the value to the public enormous.

We acknowledge while respectfully disagreeing with CRS's often-voiced concerns regarding speech and debate clause implications of publication, staff privacy, and copyright. We and others have addressed these issues at length.[14] Reports are already prepared with the possibility they will be released through a Member office or committee, by CRS to a member of the media, or by CRS to the executive branch. As online publication through non-CRS entities already exists, publication by another entity (GPO, the Clerk, etc.) would not adversely affect CRS's position. With respect to staff privacy, in some instances CRS already removes staff names from reports it believes will raise safety issues. If it so desired, it could expand that practice. Finally, as CRS reports may contain material subject to copyright by third parties, it should adopt GAO's policy of including a disclaimer.

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, for example, 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. By way of another example, while the Committee on Rules at least 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. However, in the interim, in addition to however else they are published, rules should be published in an open, non-proprietary format, even if it is 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

---

[10] "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."

[11] Legislative Branch Appropriations Bill, 1952, Hearings, pages 29-33.

[12] Legislative-Judiciary Hearings, 1954, page 11, available at http://assets.sunlightfoundation.com/policy/papers/Sen%20Leg%20approp%201954%20hearing.pdf.

[13] See http://www.gpo.gov/fdsys/bulkdata.

[14] See, e.g., Testimony Before the House Legislative Branch Appropriations Committee, FY 2012, on May 11, 2011, available at http://www.scribd.com/doc/54642878/Daniel-Schuman-Testimony-Appropriations-Subcommittee-2011-05-11

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 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.[15] 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. (It also is published every other year as a series of less-than-useful books or pocket-part updates.) In light of the House's drive to broaden the availability of documents in machine-readable formats, this issue is ripe for resolution. At a minimum, publication of either the XML source or the HTML pages would address many of our 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.

House Expenditure Reports

The quarterly House Expenditure Reports contain all spending by the House of Representatives and are currently published online as a PDF. They should be published as data files, such as CSV or XLSX, to allow for the public to easily analyze the information. The online publication that started in 2009 was a significant step forward, but the data should be available in a more flexible format.

We appreciate your attention to these issues. If you would like 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 202-733-8976 or zgraves@rstreet.org.

Sincerely yours,

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

Congressional Data Coalition
Data Transparency Coalition
Demand Progress
GovTrack.us
LegisWorks.org
OpenTheGovernment.org
R Street Institute
Sunlight Foundation
The OpenGov Foundation