Comments for the Record

from Daniel Schuman
Policy Counsel of the Sunlight Foundation
Director of the Advisory Committee on Transparency

for the Committee on Appropriations
Subcommittee on Legislative Branch
United States House of Representatives

on the Budget for
the Library of Congress and
the Government Printing Office,
regarding bulk access to
THOMAS legislative information

February 6, 2012

Comments of the Sunlight Foundation

February 6, 2012

Chairman Crenshaw, Ranking Member Honda, and members of the Committee, thank you for the opportunity to submit comments on the budget for the Library of Congress and the Government Printing Office.

I am the Policy Counsel for the Sunlight Foundation, a non-partisan non-profit dedicated to using the power of the Internet to increase government openness and transparency, and Director of the Advisory Committee on Transparency, a project of the Sunlight Foundation that brings together organizations from across the political spectrum in support of the Congressional Transparency Caucus' mission of educating policymakers on transparency issues.

Today's comments are focused on the failure of the Library of Congress to meet Congress's charge to "report on the feasibility" of "enhancing public access to legislative documents, bill status, summary information, and other legislative data through more direct methods such as bulk data downloads and other means of no-charge digital access to legislative databases."[1] Four years have elapsed since the Library said it "would look into the issue" in response to congressional prompting;[2] three years have passed since appropriators directed the Library to undertake a study;[3] and I testified about ongoing failures to make progress on bulk access to THOMAS data before this Committee last May.[4]

Providing bulk access to data means that users can download all the information contained in a database at once. By contrast, an Application Programming Interface, or API, allows computers to ask a database for specific information. THOMAS does not support either of these technologies. Instead, programmers must build tools called web scrapers that simulate a person going to each page of a website, copying that information into a database, and then trying to put those results into context. This is very hard to do automatically, particularly with large quantities of information, and the scrapers often break or take a lot of time to gather all the necessary information.

---

[1] Committee Print of the House Committee on Appropriations that accompanied the House Committee on Appropriations Omnibus Act of 2009, P.L. 111-8 (March 2009); page 1770 of the print, available at http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111_cong_house_committee_prints&docid=f:47494g.pdf (or http://1.usa.gov/lYSHd9).

[2] "Lawmakers favor outside access to legislative data," *Government Executive* (Jan. 23, 2008), available at http://www.govexec.com/oversight/2008/01/lawmakers-favor-outside-access-to-legislative-data/26148/ (or http://bit.ly/A4c5le).

[3] See the Committee Print at footnote 1.

[4] "Daniel Schuman testimony before the House Committee on Appropriations regarding CRS and THOMAS," *Sunlight Foundation* (May 11, 2011), available at http://sunlightfoundation.com/policy/documents/daniel-schuman-testimony-house-committee-appropria/ (or http://bit.ly/AsrHN4).

Recognizing this problem and the importance of public access to information, the government already provides bulk access to many datasets. The Government Printing Office, one of the entities responsible for THOMAS, has published six legislative datasets online in bulk, including the Code of Federal Regulations and the Federal Register.[5] Data.gov, which provides the public bulk access to government data, contains 3,824 "high value" data sets as of February 3, 2012,[6] with 1.5 million data downloads in the last year.[7] Compared to this veritable feast of information, THOMAS provides only a small morsel at a time.

As I mentioned earlier, there are ongoing efforts to scrape THOMAS, but these methods are prone to error, onerous, slow, and fragile. Even so, the scraped data is gathered and used by websites like GovTrack.us and OpenCongress.org, which increase the audience for congressional information by providing better user interfaces and adding important context. This data is often used on mobile platforms, too. The Sunlight Foundation's Congress app for the Android[8] smart-phone has been downloaded over 400,000 times.

A variety of non-government developers are extending the reach and value of legislative information. Much important information is being made available at no cost to the public. Its dissemination improves everyone's awareness of what's going on in Congress. These private sector efforts are necessarily limited because of the difficulty of getting the data from the Library and GPO in the first place. Legislative support agencies should recognize that aiding non-governmental efforts to disseminate legislative information is a crucial component of their public service mission.[9]

Congress has already recognized the importance of sharing legislative data

---

[5] The GPO's bulk data website is available at http://www.gpo.gov/fdsys/bulkdata (or http://1.usa.gov/kukxRG).

[6] Of the 391,428 data sets, 3,824 are so-called "high value" data sets, while the vast majority is Geodata. See http://www.data.gov/metric (accessed 2/4/2012).

[7] Data.gov monthly download trends, available at http://www.data.gov/metric/visitorstats/monthlyredirecttrend (accessed 2/4/2012).

[8] The Android app is provided free of charge at https://market.android.com/details?id=com.sunlightlabs.android.congress&hl=en (or http://bit.ly/wMq6ZE).

[9] Then-Public Printer Bob Tapella wrote in March 2009 that, in response to Congress' request to discuss access to bulk data, "a Legislative branch task force has been assembled consistent of representatives from the offices of the Secretary of the Senate, the Clerk of the House, the Library of Congress, Congressional Research, the Law Library of Congress, and GPO. This task force has already met and is working to develop a position on access to bulk data." See "Response to James Jacob's FreeGovInfo Comments," *FreeGovInfo* (April 13, 2009), available at http://freegovinfo.info/node/2509#comment-26446. Later that year, he stated "I also believe that the Federal Government has an obligation to provide complete legal and regulatory information online in an electronic format that is fully usable by the American people free of charge." See "Remarks from the Public Printer of the United States Robert C. Tapella," FDLP (October 19, 2001), available at http://www.fdlp.gov/home/repository/doc_view/1089-public-printer-remarks (or http://bit.ly/xo9n22).

broadly. In 2009, Congress adopted a forward-thinking approach that would have required an examination of granting the American people access the entirety of the legislative archives at once – via "bulk" access – in its explanatory statement accompanying the Omnibus Appropriations Act of 2009.[10] It said:

> Public Access to Legislative Data.--**There is support for enhancing public access to legislative documents, bill status, summary information, and other legislative data through more direct methods such as bulk data downloads and other means of no-charge digital access to legislative databases**. The Library of Congress, Congressional Research Service, and Government Printing Office and the appropriate entities of the House of Representatives are directed to prepare a report on the feasibility of providing advanced search capabilities. This report is to be provided to the Committees on Appropriations of the House and Senate within 120 days of the release of Legislative Information System 2.0.

The House had initially wanted to go further, proposing a report from the Library of Congress within 90 days of enactment of the 2009 legislation,[11] but the requirement was changed to no later than 120 days after the release of LIS 2.0.[12] A report was anticipated to be released during the first part of 2009.[13] Three years later, the Library has apparently ignored Congress' mandate.

---

[10] Explanatory statement available at http://bit.ly/kEiQeN (or http://www.opencongress.org/wiki/THOMAS_bulk_data_access#Policy_Documents_and_Gov.27t_Resources).

[11] Description of the original proposal is available in "Legislative Database recommendation makes it to House Leg Branch Appropriations markup," *Open House Project* (July 14, 2008), available at http://www.theopenhouseproject.com/2008/07/14/legislative-databases-recommendation-makes-it-to-house-leg-branch-appropriations-markup/ (or http://bit.ly/w4ZhcW). The original text of the proposal:

> The Committee believes that the public should have improved access to legislative information through more advanced search capabilities such as those available through the Library of Congress' Legislative Information System. The Committee also supports enhancing public access to legislative documents, bill status, summary information, and other legislative data, **through more direct methods such as bulk data downloads** and other means of no-charge digital access to legislative databases. The Committee requests that the Library and Government Printing Office **report on the progress towards these goals within 90 days of enactment of this Act**. (emphasis added)

[12] It is unknown whether LIS 2.0 is still an ongoing project within the Library. LIS 2.0 was mentioned in the Congressional Research Services' *Annual Report for Fiscal Year 2009*, but was not mentioned in the 2010 report. The 2011 report has not yet been released to the public, and the 2009 report has been removed from CRS' website. However, the 2009 report is available from the Sunlight Foundation at http://assets.sunlightfoundation.com.s3.amazonaws.com/policy/papers/crs09_annrpt.pdf (or http://bit.ly/jFTzsz), and the 2010 report is available at http://assets.sunlightfoundation.com.s3.amazonaws.com/policy/papers/CRS/crs10_annrpt.pdf (or http://bit.ly/w2STqq).

[13] See "Lawmakers favor outside access to legislative data," at footnote 2.

Movement has been so slow that the House of Representatives has been able to build and implement a system within a year that makes many primary House documents available online in bulk,[14] with more information to go online soon. A major focus of the exemplary House Legislative Data and Transparency Conference, hosted by the Committee on House Administration on February 2, 2012, was the importance of bulk access to legislative information.[15] The Library of Congress and GPO are being left in the dust. They must be prompted to act.

Times have changed since the Committee's original unheeded directive, and we request your renewed attention. We urge the committee to direct the Library of Congress, the Government Printing Office, and the Congressional Research Service – or the agencies that now have responsibility for THOMAS – to provide bulk access to legislative documents, bill status, summary information, and other legislative data within 120 days. In addition, we ask for the immediate creation of an advisory committee composed of members of these agencies and members of the public that regularly meets to address the public's need for public access to this information and the means by which it is provided. Only sustained attention can ensure that we finally make progress.

I appreciate the opportunity to draw your attention to this matter, and I welcome any questions that you may have. I can be contacted at dschuman@sunlightfoundation.com or 202-742-1520 x 273.

---

[14] "House Launches Transparency Portal," *Sunlight Foundation* (Jan. 13, 2012), available at http://sunlightfoundation.com/blog/2012/01/13/house-launches-transparency-portal/.

[15] The conference website is http://cha.house.gov/about/contact-us/legislative-data-conference (or http://1.usa.gov/yCYNcD).