ON PUBLIC ACCESS TO LEGISLATIVE INFORMATION: RECOMMENDATIONS TO THE BULK DATA TASK FORCE

August 24, 2012

These recommendations were collaboratively authored by Tom Bruce of Cornell University's Legal Information Institute; Eric Mill of the Sunlight Foundation; Daniel Schuman of the Sunlight Foundation; Josh Tauberer of GovTrack.us; and John Wonderlich of the Sunlight Foundation. For more information, contact Daniel Schuman at 202-742-1520 x 273 or dschuman@sunlightfoundation.com.

TABLE OF CONTENTS

INTRODUCTION

The Library of Congress' launch of the website THOMAS was a milestone for transparency in 1995. The Internet has changed dramatically since then, growing from a web of static pages to a web of pages and data from which information can be downloaded and integrated into a variety of customized information resources. What it means to be on the Internet today involves not just creating a website to be browsed, but supplementing it with authoritative, structured data that facilitates the efficient reuse of information. We recommend that the House embrace structured data by publishing legislative status and other information to the Internet not only as it is now, but also in structured data formats.

This recommendation is not new. A coalition of organizations came together in May 2007 to issue the report *Congressional Information & the Internet*, which made a virtually identical declaration.[1] What has changed is that this goal is now the official policy of the leadership of the House of Representatives, who pledged to "provide bulk access to legislative information to the American people without further delay."[2]

The purpose of this report is to provide recommendations to a task force established by House leadership on how to make bulk access a reality. It specifically addresses the issues raised in the committee report accompanying the House's Legislative Branch Appropriations Bill for FY 2013.[3]

[1] *The Open House Project Recommendation Report: Congressional Information & the Internet: A collaborative Examination of the House of Representatives and Internet Technology* (May 8, 2007), available at http://assets.sunlightfoundation.com.s3.amazonaws.com/policy/papers/Open_House_Project_Report.pdf

[2] *House Leaders Back Bulk Access to Legislative Information*, Speaker John Boehner (June 6, 2012), available at http://www.speaker.gov/press-release/house-leaders-back-bulk-access-legislative-information or http://1.usa.gov/Lm7yYx.

[3] The relevant text from the committee report is included in the Appendix.

THE UNMET NEED FOR LEGISLATIVE INFORMATION

Legislative information has a wide impact. The Pew Research Center's 2010 *Government Online* report found that one in five adults who use the Internet had downloaded or read legislation during the past year.[4] Millions of Americans have historically relied on THOMAS, but over the last decade websites created in the private and nonprofit sectors have surpassed THOMAS as the go-to source for legislative information. Nearly twice as many people rely on GovTrack, OpenCongress, and other sites than on THOMAS.[5] This is a healthy development. Third-parties can contextualize information in innovative ways that are beyond the current abilities[6] and scope[7] of government websites. These services depend on the ability to collect legislative information, and to do so affordably.

Currently, non-governmental web services have no choice but to rely on brittle programs to harvest information from THOMAS's complex website. This harvesting is imperfect, expensive, and time consuming. Congress's adoption of bulk access would resolve these difficulties, in essence making the entire legislative database available for download. Doing so would ease the way for third parties to build even more innovative new tools and would ensure that Americans have the most accurate information at their fingertips.

Legislative offices also have been suffering from a lack of access to their own records. House staff regularly rely on the websites mentioned above for their research. Dozens of House Member websites and DemCom (the intranet for House Democratic staff) draw on legislative information compiled by GovTrack and POPVOX.[8] There is an internal need for bulk legislative data as well.

As a long-term goal, we believe that all official artifacts of the legislative process should be available online, in real time, as structured data that is capable of being downloaded in bulk. This includes legislative text as it moves through the process; amendments; plenary, committee, and subcommittee votes; legislative status information; hearing and markup transcripts as well as video and audio from those proceedings; committee and conference reports; documents submitted for the record; and the like. We are encouraged by the recent progress on Docs.House.Gov in making these long term goals a reality.

As a starting point, all legislative information currently published on THOMAS should be available online, in real time, as structured data that is capable of being downloaded in bulk. This

[4] *Government Online*, Pew Internet and American Life Project (April 27, 2010), available at http://pewinternet.org/Reports/2010/Government-Online.aspx or http://bit.ly/b4NcvV.

[5] Over the last six months, just GovTrack and its data partners alone have been used by 5–10 million individuals.

[6] For example, GovTrack allows users to automatically redline different iterations of the same legislation. The website Scout allows users to receive automatic alerts as legislation with particular keywords is introduced or moves through the process.

[7] For example, it is beyond the scope of THOMAS to tie in Statements of Administration Policy with the legislation they refer to. Similarly, many third party websites allow users to draw upon legislative data to customize emails to their elected representatives.

[8] *Whip Hoyer Announces House Democrats' Adoption of New Online Tool to Hear From Citizens and Organizations*, Office of the Democratic Whip Steny Hoyer, available at http://www.democraticwhip.gov/content/whip-hoyer-announces-house-democrats-adoption-new-online-tool-hear-citizens-and-organization or http://1.usa.gov/PmXpeN.

includes the full legislative text, committee reports, and bill metadata such as bill summaries, status of bills, and information on co-sponsors. While some steps have been taken in this direction, there is a lot further to go.

UNDERSTANDING BULK DATA AND STRUCTURED DATA

Real transparency can only be achieved if the public has the ability to analyze information about government activity. As datasets become larger and complex, meaningful analysis depends upon the help of computers to process records. As powerful as computers are, they don't work well with data in just any format. Data must be organized -- structured -- so that computers can make sense of it.[9] Just as spreadsheets make it possible for analysts to sum, average, and chart numbers, structured data makes it possible for analysts to search, sort, and transform any sort of data.

The House of Representatives already uses structured data for many of its operations.[10] While the THOMAS and LIS websites do not publish data in a structure that supports computer-assisted analysis, they draw their information from a comprehensive database of structured data pulled together from the House, Senate, and legislative support agencies.[11]

For structured data to be useful to the public, there must be a way to access that information. While websites like THOMAS are readable by humans, they are largely incomprehensible to computers. To resolve this problem, technologists provide data for computers either in "bulk" or via "APIs." Bulk access means that the entire dataset is provided in response to an electronic request in a computer-friendly format, whereas with an API, a single data element is provided in response to an electronic request. It's the difference between giving someone an encyclopedia versus looking up a particular entry. While each method has its merits, bulk access is the preferred way to make legislative data available to the public. It reduces the burden on the provider of information while maximizing the possible ways information can be used.[12]

There are many techniques for implementing bulk, structured data, such as the XML format, CSV spreadsheet files, FTP sites, and so on. Thus there are many data models, file formats, and distribution methods that meet the description of bulk, structured data. Congress and its legislative support agencies have already demonstrated many successful uses of XML

[9] There are other requirements as well. For an analytical framework to evaluate the openness of government information, see *Ten Principles for Opening Up Government Information*, The Sunlight Foundation (August 11, 2010), available at http://sunlightfoundation.com/policy/documents/ten-open-data-principles/ or http://bit.ly/bWAJ6A; and *Data Quality: Precision, Accuracy, and Cost* in *Open Government Data: The Book*, Josh Tauberer (April 2012), available at http://opengovdata.io.

[10] Bill and resolution text are now structured data throughout their entire operational life-cycle, from the drafting process through their publication by the Government Printing Office. The new docs.house.gov website run by the Clerk's office publishes the week ahead schedule as structured data. Roll call votes, the United States Code, the Code of Federal Regulations, and many other documents have long been published as structured data by the House Clerk, the Government Printing Office, the Office of Law Revision Counsel, and other legislative offices.

[11] This March 2008 memorandum from the Library of Congress to the Committee on House Administration, entitled *Availability of THOMAS Data*, discusses what would be required to make the underlying raw THOMAS data available to the public in the structured data format known as XML. Available in the Appendix, or at http://www.scribd.com/doc/94063191/Library-of-Congress-letter-to-Committee-on-House-Administration-on-THOMAS or http://scr.bi/Kcxx1P.

[12] For more, see *Publishing Open Data – Do you really need an API?*, Peter Kranz, available at http://www.peterkrantz.com/2012/publishing-open-data-api-design/ or http://bit.ly/GB4cyl, and *Government: Do You Really Need An API*, Sunlight Foundation (March 21, 2012), available at http://sunlightlabs.com/blog/2012/government-do-you-really-need-an-api/ or http://bit.ly/GEj3T4.

throughout the legislative process,[13] as well as developed standards[14] and established a successful coordinating body between the two Houses in the form of the XML Working Group. XML is also the standard used by other legislative bodies.[15] XML's adoption within Congress and in other legislatures combined with its inherent structure makes it a particularly suitable format for Congress to employ to make its legislative information available to the public.

[13] For example, 99% of legislation is drafted in XML, roll call votes are available in XML, and the metadata behind THOMAS is kept in an XML database.

[14] See http://xml.house.gov/ and *Standards for the Electronic Posting of House and Committee Documents & Data*, Committee on House Administration, available at http://cha.house.gov/sites/republicans.cha.house.gov/files/documents/hearing_docs/2011_12_16_posting_standards.pdf or http://bit.ly/vyiRdV.

[15] See, e.g. Akoma Ntoso website, available at http://www.akomantoso.org/; the UK government's use of XML, described at http://www.opsi.gov.uk/legislation-api/developer/formats/xml; Brazil's use of XML, described at http://blog.law.cornell.edu/voxpop/2010/10/15/lexml-brazil-project/.

THE ROLE OF AUTHENTICITY

Discussions on public access to data are occasionally obscured by an ill-defined requirement that data be "authentic." At the heart of the problem is confusion about what level of authenticity is practically necessary, and why. We recognize the need for the House to publish documents that are *accurate,* in the sense of being true to the form of the document created by the issuing body. And for some purposes, documents need to carry some quality of *authority* or *officialness*, usually in circumstances bound up with using documents in official contexts, such as for legal proceedings, where an adjudicator must know the provenance of information.

PDFs with official-looking seals are comforting to some because they remind us of the fixity of print, but there are other technologies that work as well or better, and may be more practical to apply. For example, the Government Printing Office uses cryptographic digital signatures to provide authenticity to both PDF and XML metadata files.[16] A digital signature is the electronic equivalent of a fingerprint. In terms of the ability to publish files meeting standards of integrity and authenticity, PDF and XML are equivalent.

But the advantages of XML over PDF in other areas are as distinct as night and day. XML is designed to be computer-readable, which as we noted previously is crucial if the reader is going to be able to make use of large and complex documents. PDF, on the other hand, is designed only for human readability.[17] While it is a trivial task to turn computer-readable XML files into human-readable PDFs, it is very difficult to turn PDFs into XML.

Most discussions of authenticity focus on the prospective authentication of whole documents, but the need for accuracy and verification is actually much broader. Digital text is inherently reusable and recombinant at granularities much smaller than for print. Imagine, for example, that we want to create an online training manual that contains a PDF of the latest version of a small section of the US Code. It would be impractical to build the technology for that embedded text to carry a seal telling us that it's accurate. It would be cumbersome, unnecessary, and entail enormous expense both to create the seal and to verify. Guaranteeing authenticity and a high level of integrity may only be necessary in limited circumstances, and otherwise entail significant cost without commensurate benefits. [18]

As mentioned above, a large community makes use of legislative information scraped from THOMAS by GovTrack. To the extent this republication is imperfect because of the way GovTrack must gather information from THOMAS, bulk access would address those issues. Specifically, it would allow users to verify that the information they are using is accurate and

[16] As GPO notes in its report *Authenticity of Electronic Federal Government Publications* (June 13, 2011), "The publication of the cryptographic hash values in the PREMIS metadata file, and the way FDsys structures its public URLs, makes it possible for machines to crawl and use this information to determine content integrity in bulk." Available at http://www.gpo.gov/pdfs/authentication/authenticationwhitepaper2011.pdf or http://1.usa.gov/jGVanL.

[17] See *Adobe is Bad for Government*, Sunlight Foundation (October 28, 2009), available at http://sunlightlabs.com/blog/2009/adobe-bad-open-government/ or http://bit.ly/1kTZg1.

[18] There is, so far as we know, no evidence that any altered official text has ever been offered or accepted in any legal setting, and it seems that any attempt to do so would be quickly and easily detected. The danger of deliberate forgery of legislative information seems no greater than that of inadvertent use of legal information that has become stale or superseded -- a danger that is much greater with print.

would speed up its delivery. (The question of whether the information is "official" has not impeded GovTrack's millions of users.) The status quo, where the public must rely on scraped information that is unverifiable, poses a comparatively greater burden on everyone.

While the information published on THOMAS could be transmitted in a format that is capable of authentication, that is not the current practice. For instance, THOMAS *does not* use the HTTPS protocol to ensure the integrity of its information while in transit from its servers to the end user.[19] There has been significant public outcry over the lack of bulk access to structured data,[20] but there has not been a similar public alarm regarding issues of authenticity. To the extent the issue has been raised, it has already been addressed by a model that could be readily and quickly applied to releases of new legislative information.

[19] HTTPS is the encryption protocol initially used by bank websites but is now widely in use throughout the Web. For instance, Facebook uses HTTPS.

[20] See, e.g., *Thirty Organizations Call for Bulk Access to THOMAS Data*, Sunlight Foundation (April 10, 2012), available at http://sunlightfoundation.com/blog/2012/04/10/improve-public-access-to-legislative-information/ or http://bit.ly/HEnUc2.

BUDGETARY IMPACT

Compared to other technological approaches to enhancing access to legislative information, providing access to data in bulk will deliver the best bang for the buck. Providing bulk data does not involve creating a fancy website, hiring expensive mobile developers, or requisitioning vast new infrastructure. And yet, based on our experience over the last decade, it has the furthest reach.

Bill status and summary information is already stored by the Library of Congress in an XML format. Implementing public bulk access is essentially a matter of copying those files to a public location, such as an FTP server[21] or, better, an Rsync server. (There are no privacy, security, or intellectual property concerns with providing public access to the contents of the files the substance of which are already publicly available through the THOMAS website, albeit in difficult-to-use forms.)

In addition to making the files available, the House should write documentation so that the format of these files can be understood by the analysts who will access the files.

The House report accompanying the Legislative Branch Appropriations Bill raised the specter of whether a new issue might arise -- that the House may now need to "confirm or invalidate third party analyses of legislative data based on bulk downloads in XML." This is unlikely. As much legislative data is available in bulk from third parties, the House should already be receiving these calls, and thus more reliable data would likely quell inquiries concerning validation.

It would be more appropriate to budget for a process to confirm or invalidate errors in the House data itself. The Library of Congress regularly updates THOMAS with corrections. With greater exposure to the data, data users will expect to be able to report errors and to see those errors corrected in a timely way, improving reliability for everyone.

Based upon our experiences in providing bulk access to comparable legislative information to other members of the public, we can make the following estimates regarding the budgetary impact of a bulk data project.

We estimate that a minimal preparation of the data files, the creation of a public access point, and the writing of documentation for data users will take no more than 200 hours of a skilled developer and 200 hours of a House or Library staff member with a thorough understanding of the format of the existing data files and systems. Some of this work may already have been done. We encourage the task force to consider a solution that is more than minimal, however. Additional staff time and support for data preparation would be used to ensure that the data files and documentation are clean, highly normalized, clear, and presented to the public in a manner that respects the dignity and importance of the openness of the legislative process.

Ongoing maintenance of the infrastructure involves both human labor (such as systems administration) and systems infrastructure. Based on typical systems administration

[21] See the March 2008 memo from the Library of Congress to the Committee on House Administration on making the underlying raw THOMAS data available to the public, described *supra*.

requirements, we estimate the ongoing human labor requirement to be approximately two hours per week. Based on the total amount of data in the THOMAS database, likely usage scenarios, and current cloud services provider rates, infrastructure costs would be no more than $6,000 per year.

Estimate of Recurring Infrastructure Cost

Component	**Size**	**Unit Cost**	**Annual Cost**
Storage	100 GB	$0.10 per GB per month	$120
Server	Small	$0.08 per hour	$700
New User - Full Replication*	20 per month X 100 GB	$0.12 per GB	$2,880*
Existing User - Replication of Updated Data*	1,000 users requiring 1 GB new data per month (each)	$0.12 per GB	$1,440*
		Total:	**$5,220**

(The costs associated with the components marked with an asterisk could be deferred to the end user. In a setup where the end user covers the marginal cost of data transfer, the annual infrastructure cost to Congress is reduced to $820.)

IMPLEMENTATION

While the question of who within the legislative branch should have the day-to-day responsibility of releasing legislative information to the public is a question for leadership, we do have some thoughts as to possible approaches.

First, drawing upon the *Ten Principles for Opening Up Government Information*, datasets released by the responsible party or parties should be complete, primary, and timely.[22]

- Completeness of data refers to including the entirety of the public record on a particular subject. This includes metadata that defines and explains the raw information.
- Primary data is the original information collected or constructed by the government. It includes details on how the data was collected and the original source documents recording the collection of the data. To the extent that it is deemed important, the offices or Houses that originate information could be identified in the metadata.
- Timely data is information released as quickly as it is gathered and collected, with priority given to data whose usefulness is time sensitive. To the maximum extent possible, information should be made available to the public in real-time.

Second, in order to implement these goals, we recommend creating:

- A bulk data public access point, such as an anonymous FTP or rsync server.
- A process to copy the XML data from the Library's internal systems to the public server.
- A method for data users to determine which files have changed due to the availability of new information or corrected information, and for downloading only those changes. An rsync server, deltas, or granular files with easily accessible modification dates could all provide this functionality.
- A simple system for authenticity, such as a master list of file hashes.
- A static website describing how to access the data, defining the structure of the files, and, going forward, documenting changes in the implementation of this project.
- Guidelines for future changes to the data format.

In addition, the XML format should:

- Make use of existing House standards, existing Library of Congress standards, and new standards being developed for Docs.House.Gov.
- Include any cross-walk tables necessary for normalization.
- Be properly encoded in Unicode.
- Be normalized, such as encoding date/time stamps in an ISO format.
- Have date/time stamps for the date of first publication and last update of each file or record.

We hasten to add that it is far more important for Congress to release information *now* than to perfect how it releases information at some far future date. It would be acceptable for Congress

[22] *Ten Open Data Principles*, the Sunlight Foundation (August 11, 2010), available at http://sunlightfoundation.com/policy/documents/ten-open-data-principles/ or http://bit.ly/bWAJ6A.

to engage in an iterative process whereby information is released (and documented) is increasingly sophisticated ways over time. We must avoid making the perfect the enemy of the good, especially at the cost of additional delays.

Third, the responsible party for this project should be the one best in a position to make it happen.

Fourth, the party or parties responsible for the publication of data should already have within their mission and experience the release of information to the public. Those institutions that do not share this public-facing orientation but do serve internal constituencies may be better off providing data to a legislative unit already geared toward working with the public.

Fifth, there should be a working group of internal and external (non-governmental) stakeholders that meets regularly. This group should discuss issues concerning the technological means by which information is released, the logistical questions on how that data is gathered, the establishment of standards, the inclusion of new datasets for public release, and other matters. It also should conduct an audit of the data produced or gathered by the different offices within the House, Senate, or legislative support agencies to identify and make recommendations regarding what other information should be released to the public.

Sixth, each party responsible for generating information should have a high-level point of contact for internal and external stakeholder communications for questions on technological, logistical, and policy levels.

Seventh, there should also be a single person or entity that is responsible for coordinating the publication process and is the public face of these efforts. This person or entity should have sufficient authority to set deadlines, oversee budgets, and make sure that the process is accountable.[23]

[23] For example, the British Parliament has a single office for IT matters, known as Parliament ICT. See http://www.parliament.uk/documents/upload/pi21annexpict.pdf or http://bit.ly/NLiwpc. The Director of Parliamentary ICT, Joan Miller, is responsible to the political leadership. A Committee on House Administration Hearing on September 27, 2006, addressed the issue of using technology to improve House operations, noting with concern that "there is no one office, no one organization that has the ability to look across all the different pieces of decision making." See http://www.gpo.gov/fdsys/pkg/CHRG-109hhrg31073/html/CHRG-109hhrg31073.htm or http://1.usa.gov/Q9E4Ka.

CONCLUSION

The stars have aligned for the 112th Congress. Leadership of the majority and minority, the vast majority of committee chairs and ranking members, influential members of both parties, key staff, many leaders in the legislative support agencies, and leading members of the public interest community are in agreement with using technology to make Congress more open and transparent. There is enough energy behind the idea of an Open House -- in support of a more transparent Congress -- that the Congress is on the brink of an historic step forward.

Just as the House leadership in 1994 seized the opportunity to create THOMAS, and thereby open a window into the legislative process, so too do we have an opportunity right now to open up the Congress to the American people in a way that resonates with the digital age.

FURTHER READING

Bulk Access to THOMAS resource page, available at
http://www.opencongress.org/wiki/THOMAS_bulk_data_access

Data Mining Meets City Hall, Leah Hoffman (2012), available at
http://cacm.acm.org/magazines/2012/6/149784-data-mining-meets-city-hall/fulltext.

Guidelines for Open Data Policies, Sunlight Foundation (2012), available at
http://sunlightfoundation.com/policy/opendata/

Making Metasausage, Tom Bruce (2012), available at
http://blog.law.cornell.edu/metasausage/

Open Government Data: The Book, Josh Tauberer (2012), available at http://opengovdata.io

Publishing Open Data: Do you really need an API?, Peter Krantz, available at
http://www.peterkrantz.com/2012/publishing-open-data-api-design/

The Open House Project Report: Congressional Information & the Internet: A Collaborative Examination of the House of Representatives and Internet Technology (May 8, 2007), available at
http://assets.sunlightfoundation.com.s3.amazonaws.com/policy/papers/Open_House_Project_Report.pdf

APPENDIX

SELECTION FROM APPROPRIATIONS COMMITTEE REPORT 112-511 ON LEGISLATIVE BRANCH APPROPRIATIONS BILL, 2013 JUNE 1, 2012[24]

During the hearings this year, the Committee heard testimony on the dissemination of congressional information products in Extensible Markup Language (XML) format. XML permits data to be reused and repurposed not only for print output but for conversion into ebooks, mobile web applications, and other forms of content delivery including data mashups and other analytical tools. The Committee has heard requests for the increased dissemination of congressional information via bulk data download from non-governmental groups supporting openness and transparency in the legislative process. While sharing these goals, the Committee is also concerned that Congress maintains the ability to ensure that its legislative data files remain intact and a trusted source once they are removed from the Government's domain to private sites.

The GPO currently ensures the authenticity of the congressional information it disseminates to the public through its Federal Digital System and the Library Congress's THOMAS system by the use of digital signature technology applied to the Portable Document Format (PDF) version of the document, which matches the printed document. The use of this technology attests that the digital version of the document has not been altered since it was authenticated and disseminated by GPO. At this time, only PDF files can be digitally signed in native format for authentication purposes. There currently is no comparable technology for the application and verification of digital signatures on XML documents. While the GPO currently provides bulk data access to information products of the Office of the Federal Register, the limitations on the authenticity and integrity of those data files are clearly spelled out in the user guide that accompanies those files on GPO's Federal Digital System.

The GPO and Congress are moving toward the use of XML as the data standard for legislative information. The House and Senate are creating bills in XML format and are moving toward creating other congressional documents in XML for input to the GPO. At this point, however, the challenge of authenticating downloads of bulk data legislative data files in XML remains unresolved, and there continues to be a range of associated questions and issues: Which Legislative Branch agency would be the provider of bulk data downloads of legislative information in XML, and how would this service be authorized. How would ‘‘House’’ information be differentiated from ‘‘Senate’’ information for the purposes of bulk data downloads in XML? What would be the impact of bulk downloads of legislative data in XML on the timeliness and authoritativeness of congressional information? What would be the estimated timeline for the development of a system of authentication for bulk data downloads of legislative information in XML? What are the projected budgetary impacts of system development and implementation, including potential costs for support that may be required by third party users of legislative bulk data sets in XML, as well as any indirect costs, such as potential requirements for Congress to confirm or invalidate third party analyses of legislative data based on bulk downloads in XML? Are there other data models or alternative that can enhance congressional openness and transparency without relying on bulk data downloads in XML?

[24] Available at http://appropriations.house.gov/uploadedfiles/crpt-112hrpt511.pdf

The Committee directs the establishment of a task force composed of staff representatives of the Library of Congress, the Congressional Research Service, the Clerk of the House, the government Printing Office, and such other congressional offices as may be necessary, to examine these and any additional issues it considers relevant and to report back to the Committee on Appropriations of the House and Senate.

House Leaders Back Bulk Access to Legislative Information
June 6, 2012[25]

WASHINGTON, DC – House Speaker John Boehner (R-OH), Majority Leader Eric Cantor (R-VA), Legislative Appropriations Subcommittee Chairman Ander Crenshaw (R-FL), and Oversight & Government Reform Committee Chairman Darrell Issa (R-CA) released the following statement today regarding House efforts to provide bulk access to legislative information:

"The coming vote on the Legislative Branch appropriations bill marks an important milestone for the House of Representatives: the moment lawmakers agree to free legislative information from the technical limits of years past and embrace a more open, more transparent, and more effective way of doing the people's business. Our goal is to provide bulk access to legislative information to the American people without further delay.

"The bill directs a task force to expedite the process of making public information available to the public. In addition to legislative branch agencies such as the Library of Congress and the Government Printing Office, the task force will include representatives of House leadership and key committees, as well as the Clerk of the House and the House Chief Administrative Officer.

"This is a big project. That's why accomplishing it rapidly and responsibly requires all those with a role in the collection and dissemination of legislative information to be at the table together. Because this effort ranks among our top priorities in the 112th Congress, we will not wait for enactment of a Legislative Branch appropriations bill but will instead direct the task force to begin its important work immediately.

"The offices involved in this project have been instrumental in using new technology to make the House more open. We pledged to make Congress more transparent and accessible, and from our efforts to provide legislation and updates in XML, to the video streaming and archiving of committee hearings, to our search for new ways to engage and serve the American people through events like last year's 'Hackathon' – and more – we're working to keep that pledge. Bulk data is the next and a very important step. We look forward to the task force's report and to beginning implementation of this project as soon as possible."

# # # # #

[25] Available at http://www.speaker.gov/press-release/house-leaders-back-bulk-access-legislative-information.

EXPLANATORY STATEMENT,
HOUSE COMMITTEE ON APPROPRIATIONS, OMNIBUS ACT, 2009
COMMITTEE PRINT OF THE HOUSE COMMITTEE ON APPROPRIATIONS
MARCH 2009[26]

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.

[26] Available on p. 1770 at http://www.gpo.gov/fdsys/pkg/CPRT-111JPRT47494/pdf/CPRT-111JPRT47494-DivisionG.pdf.

LIBRARY OF CONGRESS MEMORANDUM
MARCH 31, 2008

CONGRESSIONAL RELATIONS OFFICE
OFFICE OF THE LIBRARIAN

MEMORANDUM

DATE: MARCH 31, 2008

TO: COMMITTEE ON HOUSE ADMINISTRATION

FROM: CONGRESSIONAL RELATIONS OFFICE

SUBJECT: AVAILABILITY OF THOMAS DATA

The Committee on House Administration asked the Library to report back on what resources would be needed to make the underlying raw THOMAS data available to the public in XML, so that other sites can re-package the data in different ways without having to link back to THOMAS.

This report responds to that request, providing a suggestion as to how this can be achieved technically. The report also highlights some policy implications of making the underlying data available in this way. If you have any questions regarding the content, Donna Scheeder of the Law Library (7-8939) is the main point of contact on programmatic issues. For technical/infrastructure issues, contact Jim Gallagher of the Office of Strategic Initiatives
(7-9600).

At Congress' request, the Library (primarily through the Congressional Research Service, CRS) has been moving forward to convert data from basic ascii text to a more robust XML format. Data conversion work completed for LIS will be immediately available on THOMAS as well . Generally speaking, any updates to the LIS will provide a basis for future work on THOMAS, and will gradually minimize differences in functionality between LIS and THOMAS. The XML database, which is a part of the "LIS 2.0" Legislative Project, will be completed in approximately two months. The data will include bill metadata such as bill summaries, status of bills, and information on co-sponsors. Full text bills and committee reports are already available on GPO ACCESS, although not in XML. Once the LIS 2.0 database is completed, the resources will be available to copy the database daily into an Anonymous File Transfer Protocol [FTP] site so it is accessible to the public.

FTP, a commonly used protocol for transferring files over a network, allows those files to be copied and moved to another computer in the network regardless of which operating systems are involved, in order to be incorporated into another interface. Anonymous FTP means users do not need an account on the server, nor do they need a password to get the file. It means that anyone desiring to transfer the data could do so. This solution is currently employed by the Illinois General Assembly, http://www.ilga.gov/. "FTP Site" is a link directly off the Home Page which takes the user to files available for transfer.

Policy implications arising out of this action involve ownership of the data. Data for THOMAS comes from a variety of sources including the House, Senate, Congressional Research Service and the Government Printing Office. While the data is in the public domain and resides on a public website, it would be prudent to discuss with the data owners any effort to make the underlying data publicly available on THOMAS before acting to do so. We have informed CRS and GPO of the interest in providing this feature on THOMAS, and will work with the appropriate House and Senate officers and committee staff to ensure that this is indeed the direction we should take THOMAS.

CRS also offers (and will continue to identify and analyze) the following policy matters for the Committee's consideration as it proceeds with determining next steps:

- *Data Accuracy.* Once we have released the data, we need to ensure that we have the ability to retract or correct errors. Data held by THOMAS and LIS can be and often are corrected.

- *Data Permanence and Authentication*. The issue of permanent accessibility and authenticity of online legal and legislative resources is an emerging concern at both the state and federal level that may need to be addressed through legislative action. Legal documents such as bills, statutes and administrative codes, are being made available online and not authenticated.

In addition, the Library is currently working on other improvements to THOMAS. For example, "Legislative Handles," a new persistent URL service for creating links to legislative documents, have been introduced to both the LIS and THOMAS [http://www.congress.gov/ help/handles.html]. This makes it much easier to create permanent links to bills. The Library is also planning to introduce, on a limited pilot basis, RSS feeds to selected THOMAS data during the coming year. This will allow the Library to assess the demand for this type of feature as well as the technical and resource needs required for expansion of RSS use in the future.

Finally, efforts are underway at the Library of Congress to undertake a study of the relationship between the LIS and THOMAS that will serve as the basis for development of a strategic plan for THOMAS. This will provide a sound basis by which we can better assess the expectations of Congress and the public, and how best to meet them. The study will also include an examination of accuracy, permanence and authentication of legislative data, along with any attendant issues, risks and workload.