LEGISLATIVE BRANCH APPROPRIATIONS FOR 2014

HEARINGS

BEFORE THE

COMMITTEE ON APPROPRIATIONS

HOUSE OF REPRESENTATIVES

ONE HUNDRED THIRTEENTH CONGRESS

FIRST SESSION

SUBCOMMITTEE ON LEGISLATIVE BRANCH

RODNEY ALEXANDER, Louisiana, *Chairman*

C. W. BILL YOUNG, Florida
JEFF FORTENBERRY, Nebraska
DAVID G. VALADAO, California
ANDY HARRIS, Maryland

DEBBIE WASSERMAN SCHULTZ, Florida
JAMES P. MORAN, Virginia
SANFORD D. BISHOP, JR., Georgia

NOTE: Under Committee Rules, Mr. Rogers, as Chairman of the Full Committee, and Mrs. Lowey, as Ranking Minority Member of the Full Committee, are authorized to sit as Members of all Subcommittees.

ELIZABETH C. DAWSON, *Clerk*
JENNIFER PANONE, *Professional Staff*
CHUCK TURNER, *Professional Staff*

PART 2

FISCAL YEAR 2014 LEGISLATIVE BRANCH APPROPRIATIONS REQUESTS

Printed for the use of the Committee on Appropriations

U.S. GOVERNMENT PRINTING OFFICE
81–322 WASHINGTON : 2013

KAREN L. HAAS
CLERK

ROBERT F. REEVES
DEPUTY CLERK

H-154 THE CAPITOL

Office of the Clerk
U.S. House of Representatives
Washington, DC 20515-6601

December 31, 2012

The Honorable Ander Crenshaw
Chairman
Appropriations Subcommittee on the
Legislative Branch
HT 2 U.S. Capitol
Washington, D.C. 20515

The Honorable Michael M. Honda
Ranking Member
Appropriations Subcommittee on the
Legislative Branch
1016 Longworth House Office Building
Washington, D.C. 20515

The Honorable E. Benjamin Nelson
Chairman
Appropriations Subcommittee on the
Legislative Branch
135 Senate Dirksen Office Building
Washington, D.C. 20510

The Honorable John Hoeven
Ranking Member
Appropriations Subcommittee on the
Legislative Branch
135 Senate Dirksen Office Building
Washington, D.C. 20510

Dear Mr. Chairmen and Ranking Members:

Pursuant to H. Rept. 112-511 that accompanied H.R. 5882, Legislative Branch Appropriations Act, 2013, as passed by the House on June 8, 2012, please find the enclosed report. As you know, the House Committee Report raised several questions regarding the increased dissemination of congressional information via bulk data download from non-governmental groups supporting openness and transparency in the legislative process. The Report directed the establishment of a task force composed of staff representatives from the Library of Congress, Congressional Research Service, Office of the Clerk, Government Printing Office, and such other congressional offices as may be necessary, to examine these questions and any additional issues considered relevant. In addition, it directed the Task Force to present its findings to the House and Senate Committees on Appropriations.

The enclosed report outlines the Bulk Data Task Force findings in accordance with the Committee's direction.

With best wishes, I am

Sincerely,

Karen L. Haas

Karen L. Haas

KLH/jd
Enclosure

LEGISLATIVE BRANCH BULK DATA TASK FORCE

Report of Activities

Submitted to
Committee on House Administration
Legislative Branch Subcommittee

December 31, 2012

Table of Contents

APPENDICES

Foreword

Improvements in technology over the last decade, especially increased capacity and lower costs, offer legislatures around the world new opportunities to foster democratic values of openness and accountability. The emergence of a diverse array of digital platforms, from social media to smart phones, are enabling greater and more substantive citizen participation in the political process. This is creating a new expectation for legislatures to be more transparent and adopt digital technologies that will enable citizens to stay informed by making information available on a timely basis in machine-readable and re-usable formats.

The United States House of Representatives has been a leader and a participant at an international level in promoting transparency and open government. In November 2009, the House co-organized the World e-Parliament Conference in Washington, DC with the United Nations, the Inter-Parliamentary Union and the Global Centre for Information Communication and Technology in Parliament. This event provided a platform for over 90 delegations from legislatures around the world to exchange views on the latest trends and different means of implementing new technologies with a view to identify good practices in the areas of representation, transparency, accountability, openness and effectiveness.

At the start of the 112th Congress, the House adopted a Rules Package that identified electronic documents as a priority for the institution. In a quote from a letter to Karen L Haas, Clerk of the House, Speaker Boehner and Majority Leader Cantor stated that, "The new House Majority is dedicated to changing how our institution operates, with an emphasis on real transparency and greater accountability. Openness, once a proud tradition of the House, is again the standard". The letter went on to direct the Committee on House Administration to establish and maintain electronic data standards for the House and its committees while tasking the Clerk with ensuring the consistent public availability and utility of the House's legislative data; these standards have been adopted and implementation is well underway.

H. Rept. 112-511 that accompanied H.R. 5882 Legislative Branch Appropriations Act 2013, directed the establishment of a task force composed of staff representatives from the Library of Congress, Congressional Research Service, Office of the Clerk, Government Printing Office, and such other congressional offices as may be necessary, to examine increased dissemination of congressional information via bulk data download from non-governmental groups supporting openness and transparency in the legislative process and any additional issues considered relevant.

The Bulk Data Task Force represents the first time that all legislative entities have been brought together in a coordinated effort on transparency and openness. The Task Force seeks to build on initial progress made over the last couple of years with such implementations as HouseLive.gov, a streaming web video of live and archived House Floor proceedings; video clipping of live and archived House Floor proceedings; HouseLive for mobile devices; updates to the House Floor Summary website; and a series of three bi-partisan meetings encouraged by House Leadership as follows:

- Congressional Hackathon was held in December 2011 and organized by House Leadership this non-partisan meeting brought together developers and Program Monitoring Organizations (PMOs) to come up with strategies to address House issues. One of the discussions focused on bulk data with a plea for no more screen-scraping.

- Legislative Data and Transparency Conference held in early February 2012, organized by Leadership and the Committee on House Administration that brought together representatives from the Legislative Branch and PMO's for a discussion on data and transparency.
- Achieving Greater Transparency in Legislatures through the Use of Open Document Standards held at the end of February 2012, organized by House Leadership, the Clerk, UN/Global Centre, and IPU brought together researchers and academics, international PMOs, and parliamentary staff to discuss their use of XML.

The continuing commitment of Leadership, the Committee on House Administration, and the Committee on Appropriations combined with the availability of resources and improved coordination of Legislative Branch agency efforts will help to provide the Bulk Data Task Force with the tools necessary to meet the challenges of engaging a new generation of citizens by ensuring an accessible, open, and transparent legislative process.

1

Executive Summary

After three months of meeting weekly with representatives from all Legislative Branch entities, outside researchers, developers and academics the Bulk Data Task Force has the following recommendations with regard to bulk data, authentication of bulk data, open data and Legislative Branch transparency initiatives. The Information Technology organizations that support the various entities of the Legislative Branch have a wide audience that they serve. That audience is made up of internal users, the general public, researchers, academics and developers. Open data and transparency initiatives can satisfy multiple elements of that audience or just a single element. The focus of H. Rept. 112-511 that accompanied H.R. 5882 Legislative Branch Appropriations Act, 2013 was bulk data. The primary audience for bulk data downloads is developers, although others may also use it. Consistent with the pledge by House Leaders, the Task Force recommends that it be a priority for Legislative Branch agencies to publish legislative information in XML and provide bulk access to that data; that the XML Working Group develop and maintain standards to ensure compatibility and interoperability of all machine-readable data published by the Legislative Branch; and that the Task Force be extended to the 113th Congress to continue to coordinate, initiate and track transparency-related projects.

At the beginning of this effort the Bulk Data Task Force spent a lot of its energy discussing authentication of XML documents and bulk data files. The Government Printing Office (GPO) has researched technologies that could do this and came up empty handed. After receiving the "Recommendation to the Bulk Data Task Force" document (Appendix B) and then meeting with the authors and other outside representatives, it became obvious that the authentication requirement was something that we could accomplish. The suggestion was made to implement the model already in use by GPO and the National Archives' Office for the Federal Register (OFR) bulk data. The Federal Register User Guide (Appendix C) explains the data through a question and answer section on "Legal Status & Authenticity" and another section describing the Schema. This model was acceptable to the outside parties that attended our meeting and Task Force members. It is the Task Force's recommendation to use this model going forward with bulk data projects.

The Bulk Data Task Force has initiated three bulk data projects that support the recommendation in the first paragraph and a fourth project, a Legislative Branch "Data Dashboard", that it is currently discussing. GPO is working on bill text that is already in XML format and making it available in a bulk data file. This project will be complete in time for the start of the 113th Congress. We will be using a Bill Data User Guide to explain the data authenticity and file format. GPO will host the bulk data download on its FDSYS website. The cost of this effort, which was started in September 2012, is $75,000, with an annual operating cost of $8,000. The second bulk data project is being developed by the Library of Congress. This project will take the existing XML Bill Summary documents and put them into a bulk data file. The cost of this project for the Library of Congress is estimated at $68,000 with an on-going annual cost of $95,000, all of which requires no additional funding but does require a re-prioritization of internal resources which could affect the development of other projects such as congress.gov. This project will also begin with documents from the 113th Congress, the Library is prepared to begin this project immediately but we won't have an estimated project completion date until a more detailed project analysis can be completed. The Bill Summary Bulk Data will reside on GPO's FDSYS website. Both of these projects will have natural follow-on projects to convert documents from previous Congresses.

The third project is the Legislative Data Challenge, which will be administered by the Library of Congress in concert or conjunction with the House of Lords Library of the British Parliament. The proposal for this project invites individuals or groups within and outside of the United States to compete on two data challenges to extend the development of the international Akoma Ntoso data standard so that it can map to the respective legislative data standards currently being used by the U.S. and the U.K. Akoma Ntoso ("linked hearts" in Akan language of West Africa) defines a set of simple, technology-neutral electronic representations of parliamentary, legislative and judiciary documents for e-services in a worldwide context and provides an enabling framework for the effective exchange of "machine readable" data. The fourth project that is currently being discussed is a Legislative Branch "Data Dashboard". The concept of this project is to provide a simple intuitive interface that allows the user to more easily link to data, documents, video, artifacts, searches and services provided by the various Legislative Branch agencies. This project would also provide links to existing bulk data downloads but would be more geared toward the general public. Using web statistics from each agency we can determine the trends in what the public is most interested in accessing and make it easy to find all in one place. Right now the Task Force is still analyzing the specifics of how we can accomplish this but we think the idea has a lot of promise.

The Bulk Data Task Force recommends that it continue to meet periodically with outside groups to enhance communication, get feedback on recent projects and continue to gather ideas for new open data projects.

During the Bulk Data Task Force meetings one issue that was identified for further investigation and reporting was determining whether or not the Legislative Branch should continue its current standard of using Document Type Definitions (DTDs) to define open data document structures or make a gradual transition to the use of newer XML Schemas. DTD's were the first method used to provide a basic grammar for defining an XML document in terms of the metadata that comprise the shape of the document. An XML Schema provides this, plus a detailed way to define what the data can and cannot contain. It also provides far more control for the developer. The Bulk Data Task Force recommends asking the Legislative Branch XML Working Group to develop a "White Paper" that will compare the use of DTDs and Schemas and make a recommendation to the Bulk Data Task Force on which technology we should use going forward. If they recommend that we do transition to Schemas, the analysis should include a timeframe for the transition and estimated transition costs.

2

Background

H. Rept. 112-511 that accompanied H.R. 5882 Legislative Branch Appropriations Act, 2013 was passed by the House on June 8, 2012. The section concerning the Government Printing Office beginning on page 17 (see text below) raised several questions that were heard during testimony about requests for the increased dissemination of congressional information via bulk data download from non-governmental groups supporting openness and transparency in the legislative process. The Report also directed the establishment of a task force composed of staff representatives from the Library of Congress, Congressional Research Service, Office of the Clerk, Government Printing Office, and such other congressional offices as may be necessary, to examine these and any additional issues considered relevant. The Committee Report also directed the task force to report back to the Committee on Appropriations of the House and Senate.

> **GOVERNMENT PRINTING OFFICE**
>
> The recommendation provides $122,456,000 in budget authority for the Government Printing Office (GPO), in addition to any offsetting collections which the GPO may earn under separate authority. This amount is $3,744,000 below the fiscal year 2012 enacted level and the budget request. GPO provides publishing and dissemination services for Federal government publications to Congress, Federal agencies, Federal depository libraries, and the American public. 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 e-books, mobile web applications, and other forms of content delivery including data mash-ups 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?

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.

3

The Process

In July 2012 House Leadership began to move forward with the establishment of the Legislative Branch Bulk Data Task Force and designated the Office of the Clerk to lead the effort. After several initial meetings were held with internal House staff to better organize the effort, the first full Bulk Data Task Force group meeting was held on Thursday, September 27, 2012. A bipartisan group of representatives from the following Legislative Branch areas attended: *House Leadership, Clerk's Office, Parliamentarian, GPO, LOC/CRS, Law Revision Counsel, House Legislative Counsel, Senate, House Administration and Appropriations.* A list of the participants can be found in Appendix A. The group reviewed and agreed to the following goals, objectives and deliverables:

- Objectives
  - **a.** Increase the amount of data available for bulk data download in open XML standards.
  - **b.** Organize legislative information so that it is easily found, available and downloadable.
  - **c.** Coordinate Legislative Branch efforts on enhancing Legislative Information and track all related projects.
- Deliverables
  - **a.** Create a delivery timeline of Legislative Information projects for:
    - **i.** The remainder of the 112th Congress
    - **ii.** First session of the 113th Congress
    - **iii.** Second session of the 113th Congress
  - **b.** Produce a report to the Appropriations Committee by 12/31/2012 responding to the Committee questions contained in the House Report 112-511 as well as other Task Force findings and estimated project costs.

The Bulk Data Task Force has met regularly since that initial meeting. On Wednesday, October 17 the task Force hosted a meeting in the Capitol with representatives from Parliamentary Monitoring Organizations, Universities, American Association of Law Libraries and United Nation's Global Centre for Information Communication and Technology. The meeting focused around a discussion of a "Recommendation to the Bulk Data Task Force" document authored by five of the attendees. Although much of the discussion centered around the group's views on bulk data, the topics of open standards documents and transparency were also discussed. The Bulk Data Task Force has used these meetings to not only discuss bulk data projects but also ongoing Legislative Branch efforts to increase openness and transparency. Copies of the meeting notes are in Appendix A, and a copy of the "Recommendations Document" can be found in Appendix B.

4

Key Findings

Through the meetings with legislative branch representatives, outside entities, and the content of the "Recommendations Document" (Appendix B) submitted by Cornell University, Sunlight Foundation and GovTrack.us, the Bulk Data Task Force has made the following observations regarding bulk data, transparency and open data:

- Legislative Branch efforts on providing information and making data available for downloads appeal to a variety of audiences including internal users, the general public, researchers, academics and third party developers. Bulk data downloads appeal primarily to developers.
- Authentication of XML bulk data is not necessary in the same manner as it has been done in PDF documents. The third party developers, academics and researchers that we talked to think that it is important to verify the accuracy of bulk data but you don't need to authenticate it. Having an available user guide like the one that the Government Printing Office (GPO) and the National Archives' Office of the Federal Register (OFR) use with the Federal Register, is acceptable. The Federal Register User Guide (Appendix C) contains a question and answer section on "Legal Status & Authenticity" and a Schema Description.
- A single location for bulk data is not necessary, it is more important to know where to go to get the data.
- Internal groups in the House access third party websites to get bulk data that is created in the House but not yet made available as bulk data.
- The predictability and completeness of the data is important to be able to tell the whole story. Third party developers, academics and researchers want structured bulk data in standard open formats with predictable URLs.
- It is not necessary to engage in a large development project that would take a lot of time to try to make many different documents available for bulk data download. It would be better to begin with documents/files that we're already creating on an individual basis and

expand that to also include those documents/files in bulk. Incremental, consistent progress would be a good future course to follow.

- A new issue that was identified was whether or not to continue to move forward with formatting our XML documents using DTDs or should the Legislative Branch transition to the use of Schemas? DTDs are an older technology that was established when the Legislative Branch first started to make XML documents available to the public. Today it seems that many governments/parliaments are migrating to the use of Schemas. There would be incremental costs and a transition period for the Legislative Branch to change over to Schemas. The main question is whether or not it's the most strategic way to move in the future and, if so, how to implement that way in the most cost-effective and least disruptive manner.
- In addition to the findings listed above, the Bulk Data Task Force has also responded below to the questions identified in H. Rept. 112-511.

The challenge of authenticating downloads of bulk legislative data files in XML remains unresolved and there continues to be a range of associated questions and issues:

1. Which Legislative Branch agency would be the provider of bulk data downloads of legislative information in XML, and how would this service be authorized?

The Task Force through its meetings and discussions has determined that having a centralized bulk data download site is not necessary. What is important is to let the developers and researchers know where to get the bulk data and to provide it in an open and standardized format. Legislative Branch entities like the Office of the Clerk which has a public website, also has the infrastructure in place to provide bulk data downloads. Other entities like the Library of Congress which don't currently have that capability in its new Congress.gov website, would be able to utilize the Government Printing Office which already has the infrastructure in place with the FDSYS.gov website.

2. How would 'House' information be differentiated from 'Senate' information for the purposes of bulk data downloads in XML?

File naming conventions for bulk data would identify whether it's a House or Senate file.

3. What would be the impact of bulk downloads of legislative data in XML on the timeliness and authoritativeness of congressional information?

There would be no impact on the timeliness and authoritativeness of congressional information. We would continue to create the individual versions of documents as we do today and add them to bulk data files as they are created.

4. What would be the estimated timeline for the development of a system of authentication for bulk data downloads of legislative information in XML?

Using the National Archives' Office of the Federal Register existing model for bulk data by providing a User Guide that explains the data's "Legal Status & Authenticity" and includes a Schema Description, the timeline is immediate.

5. 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?

Third party users will have development costs for new bulk data files but not because of confirming or invalidating legislative data. They don't feel that is a requirement of them or the users of their data. They stated at our meeting with them that they have not received any questions or challenges to the bulk data that they currently provide. While the Bulk Data Task Force is committed to providing an authenticated data source, all other costs and support for personal initiatives of third parties is incumbent on those parties.

6. Are there other data models or alternatives that can enhance congressional openness and transparency without relying on bulk data downloads in XML?

The Task Force discussed the other prominent option which is developing Application Interface Programs (APIs) that can access the data. The outside groups that we met with felt that APIs could be useful when accessing individual documents, but not as productive when trying to access a lot of data such as bulk data. Bulk data downloads were identified as the primary option at this time.

5

Legislative Branch Transparency Projects

One of the Objectives adopted by the Bulk Data Task Force was to "Coordinate Legislative Branch efforts on enhancing Legislative Information and track all projects". Appendix D contains a spreadsheet of Transparency related projects currently identified in the Legislative Branch. Not all of the projects are bulk data projects but all involve making more information or data available to the public. For example the new beta Congress.gov website has a new search engine and an updated styling that makes it easier for visitors to find what they want. The new Historian website will merge facts from the old Historian website with those from the Clerk's Arts & Archives website creating a dazzling display of information and artifacts not seen on House websites before.

During the 112th Congress as a part of Leadership's open government initiative, and in accordance with the *Rules of the House of Representatives* for the 112th Congress, the House developed the first phase of the Committee Project, the docs.house.gov website. This website is a digital repository of current and archived legislation and provides timely access to approved schedules and the text of all posted legislation to be considered on the House Floor. In addition to current legislative information, past schedules and text of archived legislation are available on the site in PDF and XML formats. The second phase of the Committee Project provides access to committee documents and text of legislation being considered in committee and by the House to the congressional community and the public. Although various formats are utilized, not all have been or will be authenticated by the Government Printing Office. If available, both the PDF version and XML version of a document are posted. Documents follow the House's document naming convention. If the documents have been processed by the Government Printing Office (GPO), there may be direct links to GPO's Federal Digital System (FDSys). The new *docs.house.gov* will also provide a single common calendar of House Committee events.

The Bulk Data task Force is also planning a Legislative Data Challenge administered by the Library of Congress to engage and broaden the community of interested outside parties to extend the development of the international

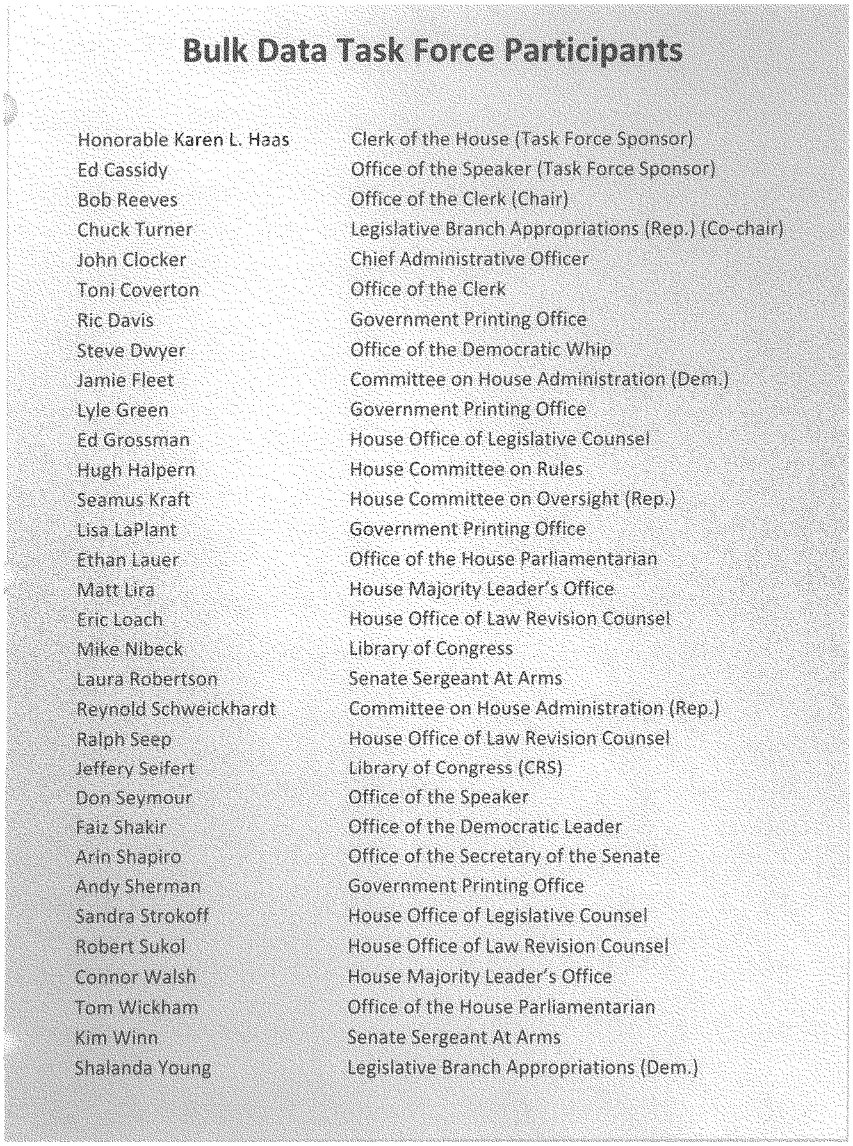
Akoma Ntoso data standard so that it can map to the respective legislative data standards currently used by the U.S. and the U.K. Another project that is currently being discussed is a Legislative Branch "Data Dashboard". The concept of this project is to provide a simple intuitive interface that allows the user to more easily link to data, documents, video, artifacts, searches and services provided by the various Legislative Branch agencies. This project would also provide links to existing bulk data downloads but would be more geared toward the general public. Using web statistics from each agency we can determine the trends in what the public is most interested in accessing and make it easy to find all in one place. Right now the Task Force is still analyzing the specifics of how we can accomplish this but we think the idea has a lot of promise.

There are several bulk data projects currently underway. The multi-phase House Modernization Project run by House Office of the Legislative Counsel (HOLC) and Office of the Law Revision Counsel (OLRC) will make the US Code available in open data standards for bulk data downloads. Two of the Bulk Data Task Force initiated projects will make bulk data files of existing Bill Text and Bill Summary open data documents available for the 113th Congress. A natural progression with two follow-on projects to convert Bill Text and Bill Summary data from previous Congresses are also available in bulk data files.

The House and the Senate are also preparing for the management of new disclosure filings required by the Stop Trading on Congressional Knowledge (STOCK) Act of 2012, which was signed into law on April 4. The Act requires financial disclosure filers to periodically report transactions in certain securities, and requires a transition to electronic reporting that will make the public disclosure information available online.

The Legislative Branch Transparency Projects spreadsheet provides the common name of the project, a brief project description, the organization or organizations responsible for the project, an estimated deployment date and a project status. It should be noted that all of these elements are based on what was known at the time the spreadsheet was printed. All project dates are subject to change based on issues discovered or changes in priorities.

CONTENTS



Bulk Data Task Force Participants

Honorable Karen L. Haas	Clerk of the House (Task Force Sponsor)
Ed Cassidy	Office of the Speaker (Task Force Sponsor)
Bob Reeves	Office of the Clerk (Chair)
Chuck Turner	Legislative Branch Appropriations (Rep.) (Co-chair)
John Clocker	Chief Administrative Officer
Toni Coverton	Office of the Clerk
Ric Davis	Government Printing Office
Steve Dwyer	Office of the Democratic Whip
Jamie Fleet	Committee on House Administration (Dem.)
Lyle Green	Government Printing Office
Ed Grossman	House Office of Legislative Counsel
Hugh Halpern	House Committee on Rules
Seamus Kraft	House Committee on Oversight (Rep.)
Lisa LaPlant	Government Printing Office
Ethan Lauer	Office of the House Parliamentarian
Matt Lira	House Majority Leader's Office
Eric Loach	House Office of Law Revision Counsel
Mike Nibeck	Library of Congress
Laura Robertson	Senate Sergeant At Arms
Reynold Schweickhardt	Committee on House Administration (Rep.)
Ralph Seep	House Office of Law Revision Counsel
Jeffery Seifert	Library of Congress (CRS)
Don Seymour	Office of the Speaker
Faiz Shakir	Office of the Democratic Leader
Arin Shapiro	Office of the Secretary of the Senate
Andy Sherman	Government Printing Office
Sandra Strokoff	House Office of Legislative Counsel
Robert Sukol	House Office of Law Revision Counsel
Connor Walsh	House Majority Leader's Office
Tom Wickham	Office of the House Parliamentarian
Kim Winn	Senate Sergeant At Arms
Shalanda Young	Legislative Branch Appropriations (Dem.)

Legislative Branch Bulk Data Task Force

Meeting Notes

Thursday, September 27, 2012 (11:00 AM - 12:00 PM)

I. Call to order

Karen Haas, Clerk of the House called to order the first regular meeting of the Legislative Branch Bulk Data Task Force at 11:10 AM on September 27, 2012 in Room HT-2, The Capitol.

II. Roll call

The following persons were present and included representatives from *House Leadership, Clerk's Office, Parliamentarian, GPO, LOC/CRS, Law Revision Counsel, House Legislative Counsel, Senate, House Administration and Appropriations*: Karen Haas, Ed Cassidy, John Clocker, Ric Davis, Steve Dwyer, Jamie Fleet, Lyle Green, Edward Grossman, Hugh Halpern, Seamus Kraft, Lisa LaPlant, Eric Loach, Mike Nibeck, Bob Reeves, Reynold Schweickhardt, Jeffrey Seifert, Arin Shapiro, Laura Robertson, Andy Sherman, Sandra Strokoff, Chuck Turner, Tom Wickam, Shalanda Young

III. Overview

Karen Haas, Clerk of the House provided an overview of the Task Force's goals which is to:

- Coordinate legislative branch efforts;
- Take the information available and decide on the best course of action to host the information;
- Increase efficiency; reduce potential redundant efforts and costs.

IV. Review of Group Objectives/Deliverables

- Objectives
    a. Increase the amount of data available for bulk data download in open XML standards.
    b. Organize legislative information so that it is easily found, available and downloadable.
    c. Coordinate Legislative Branch efforts on enhancing Legislative Information and track all related projects.
- Deliverables
    a. Create a delivery timeline of Legislative Information projects for:
        i. The remainder of the 112th Congress
        ii. First session of the 113th Congress
        iii. Second session of the 113th Congress
    b. Produce a report to the Appropriations Committee by 12/31/2012 responding to the Committee questions contained in the House

Report 112-511 as well as other Task Force findings and estimated project costs.

V. **Discussion of current bulk data/transparency projects w/n the Legislative Branch**

- The group discussed the next release of docs.house.gov in January 2013 that will provide Committee documents in PDF and XML format (when available).
- A HouseLive update that will provide video speaker search capability.
- Stock Act – Phase I implementation and Phase II which will be completed by 10/31/13.
- The recent LOC release of the beta.congress.gov website. Efforts are continuing to make all of the data currently available in Thomas also available. The infrastructure is in place to support the development API's for access to bulk data.
- GPO has expanded data downloads to make congressional bills for the last several Congresses available digitally through FDsys in PDF and XML formats.
- HIR will be launching a new mobile version of House.gov.
- Law Revision Counsel is working on a multi-year project to make the U.S. Code available in XML format to Members and the public.
- New release of the Law Revision Counsel beta website that features new search engine and is up to date through Pub. L. 112-173.
- XML votes, Hearing schedules are available on Senate.gov.
- Minority Leadership organizes talking points for Members by screen scraping data from multiple sources. They also organize Dear Colleague communications.
  - The Census Bureau website was mentioned as a great model of how bulk data can be presented to the public.
- There was a discussion of the effort to upgrade XMetal to the Windows 7 platform. The goal is to migrate bills and the amendment feature set to run in more modern environment. The goal is to have XML versions of Committee Reports available at the start of 113th Congress.
- A brief discussion was held concerning the multiple audiences that the Legislative Branch must provide information/data to:
  - Public
  - Internally to Members and staff
  - Parliamentary Monitoring organizations (PMO's)
  - Researchers and academics
- One suggestion was mentioned to potentially create a Legislative Branch data dashboard to better organize what we already have spread across multiple websites. This data dashboard could be run on all Legislative Branch websites providing the end user with seamless access to data, information and bulk data downloads.

VI. Introduction and discussion of bulk data Recommendations

- The Group was asked to review the Bulk Data Recommendations document that was emailed to them for discussion at the next meeting.
- The plan is to meet with the authors of the Bulk Data Recommendations Document in mid-October (10/17) to allow them to present their findings and explain the reasoning behind them. The group was asked to think about who should be invited to this meeting.

VII. Schedule of Next Meeting

- Next task force meeting has been scheduled for Thursday, October 4, at 11:00 AM in Room HT-2.

VIII. Adjournment

The meeting was adjourned at approximately 12 noon.

Draft meeting notes taken by: Toni Coverton, Office of the Clerk

Legislative Branch Bulk Data Task Force

Meeting Notes

Thursday, October 4, 2012 (11:00 AM - 12:00 PM)

I. Call to order

Bob Reeves called to order the regular meeting of the Legislative Branch Bulk Data Task Force at 11:02 AM on October 4, 2012 in Room HT-2, The Capitol.

II. Roll call

The following persons were present: Ed Cassidy, Ric Davis, Steve Dwyer, Lyle Green, Ed Grossman, Hugh Halpern, Seamus Kraft, Lisa LaPlant, Ethan Lauer, Mike Nibeck, Bob Reeves, Laura Robertson, Reynold Schweickhardt, Ralph Seep, Jeffrey Seifert, Arin Shapiro, Andy Sherman, Tom Ullrich, Connor Walsh, Kim Winn and Shalanda Young.

III. Approval of Minutes from Last Meeting

Bob Reeves e-mailed the minutes from the last meeting to the group. A few revisions were received and subsequently made. The updated version was approved as read.

IV. Discussion of Recommendations to the Bulk Data Task Force Document

The Task Force agreed to meet with the authors to discuss their recommendations. The meeting has been scheduled for October 17, 1:00-2:30 pm in HC-8. The purpose of the meeting is to allow the authors to present the rationale behind their recommendations and for others to be able to ask questions and state their views. A non-partisan announcement concerning the meeting will be made beforehand. The group walked through the Recommendations document. Many points and questions were raised during this walk through, and a summary of the highlights is provided below:

- We should obtain copy of the report submitted to House Leadership in May 2007, which was referenced in the introductory section of the Recommendations document.
- There was a brief discussion of the 2010 Pew research statistics quoted saying that, "one in five adults who used the internet had downloaded or read legislation during the past year."
- There was a big discussion on how to authenticate XML documents or even if we needed to.
    - The authors of the Recommendations document suggest that authentication shouldn't be a stumbling block in bulk data downloads.
    - Versioning could be a solution. StarPrint tools were mentioned.

  - Use of hash tags could be an option.
  - Bulk data is not authoritative; using a disclaimer such as the one GPO uses with the bulk data download of the Federal Register is another option.
  - The official paper copy is still considered the version of record.
- There was a discussion of the popularity and usage of bulk data downloads versus the use of API's.
  - API's are more geared toward single document downloads
  - The cost of creating an API versus a bulk data download file was estimated to be 10 to 1.
- A question was raised about focusing on a single data repository or does it make more sense to look at creating a data dashboard to seamlessly link to data where it currently resides?
- There was a discussion about who our audience(s) is:
  - Public
  - Internal users
  - PMO's / researchers
  - Who should make data available to the public? PMO's?
  - Are data fees acceptable?
  - There was mention of web service called Popvox, an online system that creates a direct and transparent line of communication for constituents and advocacy groups to reach the representatives who make decisions on their behalf.
- The document sited some infrastructure set up costs that appeared very low. We need to ask if these cost estimates reflected cloud based processing and storage?

Although there were many questions and concerns raised, not all of them could be captured in this context. The group was asked to submit their questions for the authors to Bob Reeves by October 12, which will be sifted for duplication, etc. The plan is to submit a comprehensive list of questions to the authors on behalf of the task force prior to the 10/17 meeting.

V. <u>Tentative Meeting Logistics</u>

- Keep meeting to an hour
- Meet same time and place

VI. <u>Update on a potential XML Project with GPO</u>

- The House asked GPO what they can do in the area of bulk data. A meeting has been set up for next week with GPO, Senate and House representatives. Bob Reeves will report back to the Task Force on the meeting next week.

VII. **Schedule of Next Meeting**

- The next task force meeting has been scheduled for Thursday, October 11 at 11:00 AM in Room HT-2. Please note that the October 17th meeting with be in lieu of the regularly scheduled Thursday meeting for that week.

VIII. **Adjournment**

The meeting was adjourned at around 11:57 am.

Draft meeting notes taken by: Toni Coverton, Office of the Clerk

Legislative Branch Bulk Data Task Force

Meeting Notes

Thursday, October 11, 2012 (11:00 AM - 12:00 PM)

I. Call to order

Bob Reeves called to order the regular meeting of the Legislative Branch Bulk Data Task Force at 11:01 AM on October 11, 2012 in Room HT-2, The Capitol.

II. Roll call

The following persons were present: Ed Cassidy, John Clocker, Steve Dwyer, Lyle Green, Lisa LaPlant, Eric Loach, Mike Nibeck, Bob Reeves, Laura Robertson, Reynold Schweickhardt, Jeffrey Seifert, Arin Shapiro, Ethan Lauer, Ed Grossman, Andy Sherman, Don Seymour, and Chuck Turner

III. Approval of Minutes from Last Meeting

Bob Reeves e-mailed the minutes from the last meeting to the group. The minutes were approved as read, but an extension to edit was offered.

IV. Presentation by GPO of Potential Bulk Data Project

The House and Senate met with GPO to go over a proposed project to convert bill text into bulk data XML. GPO walked through the proposed project with the group to get their feedback. Some of the highlights from the presentation included:

- **FDsys Bulk Data Site (screen shots; Federal Register case study)**
  - The FDsys Bulk Data Site contains XML for select Office of the Federal Register publications, which are also available in PDF and text formats on the main FDsys website.
- **Federal Register XML Files**
  - XML files are grouped by year and then by month. The resource folder in the root directory contains a User Guide that includes the legal status and authenticity of Federal Register XML files. Technical info is also supplied by GPO. All daily FR XML files are "zipped" for an entire year and made available on the FDsys Bulk Data Site.
- **Congressional Bills Bulk Data**
  - GPO has the capability to add congressional bill XML files to the existing FDsys Bulk Data Site beginning with the 113th Congress.
  - Bill XML files could be "zipped-up" for each bill type, (e.g. HR, HJRes, SJRES, etc)

- A resource folder could contain a User Guide with a statement regarding the legal status and authenticity of bill XML files.
- Quite a bit of flexibility built in to manipulate and set up data as needed.

Following the presentation, the group discussed the following:

- Possibility of putting together certain kinds of bills, e.g. HR and HJ or just keeping each bill type in a separate file.
- There was a concern that people looking for HJRES for example, are not inclined to look in HJ file.
- From a technical perspective, how hard would it be to grab multiple types of bills? What metadata is available for this purpose?
- Who determines structure of XML - this group or joint conversation between House, Senate, GPO, Library?
- We should continue to use current congressional standards when available.
- How does this link up to the bill summaries that CRS/LOC produces?
- What's data is available on GovTrack compared to what would be available through the proposed GPO project? GovTrack is considered a pseudo data source because it's data is all in one place.

Continued updates will be provided as this project progresses.

V. Federal Register Bulk Data User Guide

The group reviewed the Federal Register Bulk Data User Guide language to see if it could be a good example of something we should consider doing for the the potential "bill text" bulk data project in lieu of authentication.

VI. "Recommendations" Meeting Update

The "Recommendations" meeting will take place on October 17, 2012. Bob Reeves will moderate the meeting. The plan is to go through each section from start to finish. A list of topics/questions will be sent to the authors ahead of the meeting for preparation. The group was reminded to e-mail Bob Reeves the name(s) of anyone not invited to the meeting that they thought should also attend. Anticipated attendance will be 30-35 people.

VII. Future Meeting Topics

The group discussed the different ways that these meetings could be utilized in the future for things relating to XML data, technology, and transparency in general. Some of the ideas discussed were:

- Presentation from Leg Branch Working Group on XML information and standards.
- Presentation from Jeff Griffith, of the Global Centre for ICT in Parliament on current technology trends in International Parliaments around the world.
- There was also interest from the group to talk to state governments to see what they are doing. Many Congressman come from state legislatures. Kansas for example has a really aggressive technical plan and the New York Senate is doing some interesting things. Bob Reeves will reach out to the National Conference of State Legislatures (NCSL) to see if a presentation or discussion could be arranged.
- Presentation from Census Bureau and other similar groups who have been successful with creating downloadable bulk data files.

Every effort will be made to try to keep presentations between 30-45 minutes.

VIII. Schedule of Next Meeting

- The next task force meeting will be held on **October 17, 1:00-2:30 pm** in Room HC-8. This meeting with the be in lieu of the regularly scheduled Thursday meeting for the week.

IX. Adjournment

The meeting was adjourned at around 11:48 am.

Draft meeting notes taken by: Toni Coverton, Office of the Clerk

Legislative Branch Bulk Data Task Force

"Recommendations" Meeting Notes

Wednesday, October 17, 2012 (1:00 PM - 2:30 PM)

I. Call to order

Bob Reeves called to order the "Recommendations" meeting of the Legislative Branch Bulk Data Task Force at 1:02 PM on October 17, 2012 in Room HC-8, The Capitol.

II. Roll call

The following persons were present: Molly Bohmer, Tom Bruce*, Ed Cassidy, Gherardo Casini, John Clocker, Cliff Cohen, Ric Davis, Steve Dwyer, Sara Frug, Lyle Green, Jeff Griffith, Ed Grossman, Karen Haas, Jim Harper, Elizabeth Holland, John Joergensen, Seamus Kraft, Lisa LaPlant, Ethan Lauer, Eric Loach, Eric Mill*, Mike Nibeck, Bob Reeves, Laura Robertson, Daniel Schuman*, Reynold Schweickhardt, Jeffrey Seifert, Don Seymour, Arin Shapiro, Andy Sherman, Joshua Tauberer*, Chuck Turner, Connor Walsh, and John Wonderlich*
(**authors of Recommendations document)*

III. Approval of Minutes from Last Meeting

Bob Reeves e-mailed the minutes from the last meeting to the group. The minutes were approved as received, but an extension for edits was offered.

IV. Overview and Introductions

Bob Reeves provided an overview for the purpose of this meeting, which is to provide the authors of "Recommendations" document an opportunity to explain the rationale behind their recommendations, address the list of questions provided to them in advance of the meeting, and open the floor up for discussion following their presentation. Each person in attendance was asked to introduce themselves and give their affiliation.

V. Presentation of "Recommendations" Document

Daniel Schuman began the presentation by expressing the Authors' appreciation for having the opportunity to address the group. The fact that the meeting is taking place is important in a long series of efforts, and the result of the three other conferences held this year. The goal of the "Recommendations" document is to show how the House could provide bulk access to legislative information already being published on-line to the public. The authors walked through each section of the document and below are some of the highlights from the presentation:

- **The Unmet Need for Legislative Information**
  - The vast majority of users are looking to gain a better understanding of how Congress works. Well over 50,000 people rely on THOMAS for legislative information each day. Nearly twice as many individuals as well as government staffers are relying on GovTrack and other data communication partners to retrieve legislative information.
  - The discovery cost for this information is expensive and time consuming. Bulk data can be relevant not just to programmers, but also to individual users. Although the primary users of bulk data are programmers. THOMAS and Congress.gov are there to make data available, but there are resource limits as well as capacity limits.
  - As part of the Open States Project, Sunlight's Scout application is a tool that allows users to gather information about legislation in all 50 states in the same format.
  - Predictability of the data available is a key component.
  - There is also a need to have all the data not just a portion of it. Having all of the data can tell a different story.
  - If data is taken and published in bulk, it lowers barriers and costs by making the data available in a single place.

- **Understanding Bulk Data and Structured Data**
  - Structured data is a prerequisite for any sophisticated dataset. The desire is to have permanent URLs but at a bare minimum predictable URLs. Bulk data is more than having pieces of data; bulk data is intended to take information and present it to people all at once in the same form.
  - In the Open States Project, three of the 5 states that offer bulk data are the most easiest to work with. Minimal impact is needed to scrape information. The New York Senate has an initiative to develop APIs. In certain environments, APIs are handy.
  - If you have data in an XML format you can convert it to PDF or print it. If you only have data in PDF format it is extremely difficult to produce the XML. PDFs are primarily used for reading and viewing.
  - There is an understanding that data is spread out across the Legislative branch agencies and it is not necessary to have it all in one location. It is important to know where the data is and have common forms and a cross-walk of the data.
  - Having the Bio-Guide ID's has really helped with vote data.

- **The Role of Authenticity**
  - Authenticated PDF documents are the documents of record and the ones that are officially archived.
  - It is important to verify the accuracy of bulk data, but you don't need to authenticate bulk data. The States participating in the Open States Project are not authenticating bulk data.
  - Data access and authenticity are separate issues but go together. The use of a disclaimer could be a way to address not authenticating bulk data. A good example used was the way Federal Register bulk data is handled with a user guide containing specific information about the data.
  - There could be an opportunity for an authentication / phone home API to be developed that can check the version.
  - There have been little known problems from non-authenticated bulk data dissemination by third parties.
  - Third parties can provide feedback loops for the data being made available.
  - Endorse what is here; we need good documentation to support the technical efforts.
  - Don't want to undermine the retail side of the House use of gold standard documents.
  - Regular meetings among folks working on these issues; can be the canaries in the coalmine.
  - There are five states that are taking the lead in bulk data:
    - New Jersey uses an FTP site to make data available;
    - New Hampshire publishes a zip file full of text files;
    - California has more complete data;
    - North Carolina and New Mexico are also involved.

- **Budgetary Impact**
  - Scope is the big question because they don't know how easy/hard it is to make this data available; 3 main focuses -- Clerk, Library, and GPO.
  - Costs outlined are best guesses, but difficult to know internal constraints. Storage figures are based on Cloud Computing costs.
  - Authors happy to talk more about estimates at a later time

- **Implementation**
  - Incredible amount of value in incrementally doing things; don't get hung up on trying to satisfy things all at once.
  - Prioritize three main sources –pieces of data to get first, would be LOC, Clerk, and GPO.
  - Include LRC in list of data providers too.

- In terms of volume, most scraping is currently done on Thomas.

VI. Summary

- The Author's asked what the next steps of the Bulk Data Task Force would be?
    - Internal meetings to assess what was heard today in preparation for responding to the Appropriations Committee.
    - Continued progress with bulk data projects already in development and identification of future projects.
    - Potential continued meetings with PMO's and the public.
    - More slow but steady progress as we have made in the past couple of years.

VII. Schedule of Next Meeting

- The next task force meeting will be held on October 25, at 11:00 AM in Room HT-2.

VIII. Adjournment

The meeting was adjourned at around 2:27 PM.

Draft meeting notes taken by: Toni Coverton, Office of the Clerk

Legislative Branch Bulk Data Task Force

Meeting Notes

Thursday, October 25, 2012 (11:00 PM - 12:00 PM)

I. Call to order

Bob Reeves called to order the meeting of the Legislative Branch Bulk Data Task Force at 11:05 AM on October 25, 2012 in Cannon B-106.

II. Roll call

The following persons were present: John Clocker, Ric Davis, Lyle Green, Jeff Griffith (Global Centre), Ed Grossman, Seamus Kraft, Lisa LaPlant, Tom Wickham, Eric Loach, Mike Nibeck, Bob Reeves, Jeffrey Seifert, Don Seymour, Arin Shapiro, Andy Sherman, Rob Sukol, and Ralph Seep.

III. Approval of Minutes from Last Meeting

Bob Reeves e-mailed the minutes from the last meeting to the group. The minutes were approved as received, but an extension for edits was offered.

IV. Discussion of the "Recommendations" meeting held last week

- The group discussed the meeting held last week with the "Recommendations" document authors and some other invited guests. The consensus was that the meeting was successful for everyone that attended. It provided a good example of the commitment of the Legislative Branch to continue to work toward opening up its' documents to the public. It also gave the authors and other invited guests a vehicle to state their views and recommendations as well as get some face time with representatives from across the Legislative Branch. The items discussed are already documented in the meeting notes from the 10/17 meeting.

V. Presentation of the 2012 World e-Parliament Report

- Jeff Griffith from the UN's Global Centre for ICT in Parliaments did a presentation on the 2012 World e-Parliament Report. Jeff highlighted the process they used to gather the information, the technology trends identified and related the recent findings to the results of previous Reports. The report can be downloaded from the Global Centre website at: http://www.ictparliament.org/WePReport2012. Jeff stated that he would have some hardbound copies in the near future and would pass them on to Bob Reeves for distribution. A copy of Jeff's presentation will be submitted along with this week's meeting notes.

VI. Update on Potential Projects

- Bob Reeves discussed a couple of potential projects that are currently under consideration with the Library of Congress (LOC):
    - Bill Summary bulk data – the LOC is putting together an analysis of what it would take to do this.
    - Legislative Branch Challenge – the LOC is also working on providing the framework for a potential challenge for college students to take XML versions of House bills and create an Akoma Ntoso (international standard) version. There is also some discussion of partnering with the UK on a joint effort.

VII. Schedule of Next Meeting

- The next task force meeting will be held on November 1, at 11:00 AM in Cannon B-106.*
- Upcoming presentations are scheduled as follows:
    - 11/8 - Ed Grossman and Rob Sukol on the HOLC/OLRC Modernization Project
    - 11/15 - Kirsten Gullickson and Marsha Misenhimer on the XML Working Group and new features of XMetal 7.0

**Please note that the November meetings will be held in Cannon B-106 because of scheduled presentations.*

VIII. Adjournment

The meeting was adjourned at around 12:01 PM.

Draft meeting notes taken by: Bob Reeves, Office of the Clerk

Legislative Branch Bulk Data Task Force

Meeting Notes

Thursday, November 8, 2012 (11:00 AM - 12:00 PM)*

**Please note that the BDTF Meeting scheduled on November 1 was cancelled due to the short work week because of inclement weather.*

I. Call to Order

Bob Reeves called to order the meeting of the Legislative Branch Bulk Data Task Force at 11:03 AM on November 8, 2012 in Cannon B-106.

II. Roll Call

The following persons were present: Ric Davis, Steve Dwyer, Ed Grossman, Hugh Halpern, Lisa LaPlant, Ethan Lauer, Bob Reeves, Laura Robertson, Reynold Schweickhardt, Ralph Seep, Jeffrey Seifert, Don Seymour, Arin Shapiro, Andy Sherman, and Rob Sukol.

III. Approval of Minutes from Last Meeting

Bob Reeves e-mailed the minutes from the last meeting to the group. The minutes were approved as received, but an extension for edits was offered.

IV. Presentation on HOLC/OLRC Modernization Project

- Rob Sukol and Ed Grossman each did a presentation on the House Modernization Project, which is a joint project being undertaken by the House Office of the Legislative Counsel (HOLC) and the Office of the Law Revision Counsel (OLRC). OLRC's part of the House Modernization Project was presented in 3 stages:
  - Stage 1 – Conversion (converts U.S. Codes into free downloadable XML bulk data; most relevant to Bulk Data Task Force)
  - Stage 2 – Positive Law Codification System (restates existing laws, improves organization and removes obsolete provisions)
  - Stage 3 – Editorial Updating System (edits and updates the U.S.C. using various programs based on MicroComp locators to meets the highest standards of accuracy.
- HOLC's objectives in the Modernization Project are to improve:
  - Quality and consistency of amendment drafts to laws and bills
  - Collaboration and communications within HOLC and with clients
  - Management of workflow within HOLC
- HOLC just released a new Intra-House website last month, http://www.house.gov/legcoun/

- Timeframe to complete all 3 stages of the project will likely be towards the end of 2014. The target for Stage 1 will be completed by spring 2013.
- For more details, please refer to presentation slides.

V. Data Dashboard Ideas Discussion

- Mr. Reeves presented a mock-up data dashboard concept as a starting point. The concept included all associated logos (in alpha order) - - GPO, House, LOC, Senate with drop down categories to search for information by popularity. ***Please see attached mock up for further details.***
- The group discussed a number of thoughts that included:
    - How will the information be organized on the Data Dashboard (DDB)?
    - How to list the popular links.
    - Who will develop? Maintain? Update?
    - Develop with what technology? Something that can run on all websites.
    - Will public see same thing on each website?
    - Look at what databases people are using?
    - How to characterize goals of DDB? Who is our audience? Public? Researchers?
    - How to minimize confusion and steering focus to use one data source over another.
    - Mobile apps, web page, widget?
    - Would everyone put it on their webpage?

- Overall, everyone liked the pragmatic approach and the idea of organizing the webpage by groups. There should be one objectives page. It should also document what we've done thus far, and other projects "coming soon".
- The group was asked to continue to think about ideas and be prepared to discuss further during the next meeting.

VI. Update on Potential Projects

- Bob Reeves provided an update on the projects that are currently under consideration with the Library of Congress (LOC):
    - Bill Summary bulk data – Received proposal from the LOC; waiting for guidance.
    - Legislative Branch Challenge – the LOC is also working on providing the framework for a potential challenge to take XML versions of House documents and create Akoma Ntoso (international standard) versions. This may be a joint effort with the UK House of Lords.

VII. Schedule of Next Meeting

- The next task force meeting will be held on November 15, at 11:00 AM in Cannon B-106, and will include a presentation by Kirsten Gullickson and Marsha Misenhimer on the XML Working Group and new features of XMetal version 7.0.
- We will not have a meeting the week of Thanksgiving. The next task force meeting will be held on **November 29, at 11:00 AM in Cannon B-106**. Matt Wasniewski, House Historian, was asked to do a presentation on the new Historian's website.

VIII. Adjournment

The meeting was adjourned at around 12:02 PM.

Draft meeting notes taken by: Toni Coverton, Office of the Clerk

Legislative Branch Bulk Data Task Force

Meeting Notes

Thursday, November 15, 2012 (11:00 AM - 12:00 PM)

I. Call to Order

Bob Reeves called to order the meeting of the Legislative Branch Bulk Data Task Force at 11:01AM on November 15, 2012 in Cannon B-106.

II. Roll Call

The following persons were present: John Clocker, Steve Dwyer, Lyle Green, Ed Grossman, Kirsten Gullickson, Seamus Kraft, Matt Lira, Marsha Misenhimer, Mike Nibeck, Bob Reeves, Laura Robertson, Ralph Seep, Jeffrey Seifert, Arin Shapiro, Rob Sukol, and Connor Walsh.

III. Approval of Minutes from Last Meeting

Bob Reeves e-mailed the minutes from the last meeting to the group. The minutes were approved as received, but an extension for edits was offered.

IV. Presentation by XML Working Group

- Kirsten Gullickson and Marsha Misenhimer each did a presentation on the Legislative Branch XML Initiative, which is an on-going Working Group comprised of office representatives from both the House and Senate, as well as the LOC and GPO. The primary goal for this working group since its inception in 1996 is to produce and exchange legislative documents using the industry-based standards approach of the Extensible Markup Language (XML).
- Marsha Misenhimer, Secretary of the Senate's Office, began the presentation with highlighting the Working Group's background and outlining some of its activities and responsibilities which included:
    - Creating and managing the Common Tag Library and DTDs (document type definition) and managing issues involving exchanging and drafting legislative documents using these DTDs.
    - Collaborating to find solutions to issues, developing joint tools, and advising respective oversight organizations.
    - Meet periodically with different vendors and other organizations working on XML software and XML projects to discuss the latest tools and features.
- Kirsten Gullickson, Office of the Clerk's Legislative Computer Services, focused on presenting XMetal features.

- The Working Group's latest projects include:
    - Creating DTDs for US Code and Committee Reports
    - Partnering with GPO to have a shared tool to record committee votes.
    - Working with Rules and the Science Committee to produce resolutions in XML.
- One of the issues raised concerned making a decision to transition from the use of DTD's to schema's. This will be a topic of further discussion.
- After the presentation, there was a brief discussion of the process for moving to XMetal 7 in a Windows 7 environment. The Working Group will continue to maintain older versions for a short time, until everyone has switched over to the same operating system.
- For more information, please see full copy of the presentation.

V. Data Dashboard Ideas Discussion

- The bulk data group was not able to discuss data dashboard ideas as time ran out. Bob Reeves pushed the discussions to the next meeting.

VI. Update on Potential Projects

- Bob Reeves provided an update on the projects that are currently under consideration:
    - LOC provided a proposal for the Legislative Branch Challenge. The idea is to take Legislative Branch documents and make them available in an international standard version. The international challenge would run from January to June. More information to come on this.
    - The Parliamentarians and the Rules Committee have developed a data disclaimer that would be added to docs.house.gov with the roll out of Phase 2 of the Committee Repository Project in January 2013. The American Association of Law Library's had requested that we consider a disclaimer earlier this year. Bob read the proposed disclaimer language to group and will send it with the meeting minutes.

VII. Schedule of Next Meeting

- The next task force meeting will be held on November 29 at 11:00 AM in Cannon B-106. Matt Wasniewski, House Historian, was asked to do a presentation on the new Historian's website. We hope everyone has a happy and safe Thanksgiving Holiday!

VIII. Adjournment

The meeting was adjourned at 12:00 PM.

Draft meeting notes taken by: Toni Coverton, Office of the Clerk

Legislative Branch Bulk Data Task Force

Meeting Notes

Thursday, November 29, 2012 (11:00 AM - 12:00 PM)

I. Call to Order

Bob Reeves called to order the meeting of the Legislative Branch Bulk Data Task Force at 11:02AM on November 29, 2012 in Cannon B-106.

II. Roll Call

The following persons were present: Ric Davis, Farar Elliott, Lyle Green, Ed Grossman, Lisa LaPlant, Ethan Lauer, Dal Multani, Mike Nibeck, Laura O'Hara, Bob Reeves, Ralph Seep, Jeffrey Seifert, Don Seymour, Arin Shapiro, Sandy Strokoff, Rob Sukol and Matt Wasniewski.

III. Approval of Minutes from Last Meeting

Bob Reeves e-mailed the minutes from the last meeting to the group. The minutes were approved as received, but an extension for edits was offered.

IV. Presentation by the House Historian

- Matt Wasniewski, House Historian, Farar Elliott, Chief and Curator of Art and Archives and Laura O'Hara also of Art and Archives did a joint presentation on the exciting features of the new Historian's website. Dal Multani of Legislative Computer Systems provided support for technical questions. The new website is intuitive, user friendly and brings together three areas of House heritage to one site. Some of the highlighted features will include:
  - Roughly 600 historical highlights and approximately 1,000 objects from the House collection
  - Congressional profiles
  - Oral Histories on profound events such as 9/11 and the Civil Rights
  - Weekly blogs on unique stories (first step to introduce social media onto website; other social media formats such as Twitter will soon follow.)
  - Keyword searches and clickable "blue links" to learn more about a particular feature
  - Finding Aids (first time this features has been available on-line)
  - Geographical maps of the House
  - Expansion capability
- Attendees were pretty excited about the website, and looked forward to introducing it within their own divisions. Although the addition of committee

rosters will be added to the site over time, there was some discussion about the possibility of adding Committee history as well.

- A soft launch of the new website is scheduled the week of December 17. Bob Reeves will notify the bulk data group via e-mail on the actual launch date. So please stay tuned!

V. **Data Dashboard Ideas Discussion**

- The bulk data group was not able to discuss data dashboard ideas as time ran out. Bob Reeves gave the group homework to focus on labeling, call categories and any other grouping information not captured on the sample web page. Ideas/suggestions should be sent to Toni Coverton. Discussions will hopefully resume during the next meeting.

VI. **Update on Potential Projects**

- An update on the projects that are currently under consideration was provided:
  - Bob Reeves asked the LOC for a possible date to present the proposed Legislative Branch Challenge to the bulk data group. No confirmed date as yet.
  - He also asked Kirsten Gullickson (LCS) to do a presentation to the group on the new features that are part of the Phase 2 roll-out of docs.house.gov.
  - Couple of new things happening in the Clerk's Office:
    - The Excel Spreadsheet and ASCII Text Files available for download on the Clerk's website under the Member Information tab / Official Lists has been updated to include the Member's 112th Congress State/District and their BioGuide ID. This update was based upon a suggestion made by Reynold Schweickhardt after a discussion with Jim Harper of CATO as something that was desirable and also happened to be easy enough to accomplish. A good example of low hanging fruit.
    - Bob Reeves e-mailed the screenshot of the expanded spreadsheet and ASCII text files to the bulk data group on 11/30.
  - In the coming weeks Floor Summaries will be available in bulk data XML format by Congress/by Session in addition to the single day downloads.

VII. **Schedule of Next Meeting**

The next task force meeting will be held on December 6 at 11:00 AM in Cannon B-106.

VIII. **Adjournment**

The meeting was adjourned at 11:58 AM.

Draft meeting notes taken by: Toni Coverton, Office of the Clerk

Legislative Branch Bulk Data Task Force

Meeting Notes

Thursday, December 6, 2012 (11:00 AM - 12:00 PM)

I. Call to Order

Bob Reeves called to order the meeting of the Legislative Branch Bulk Data Task Force at 11:02AM on December 6, 2012 in Cannon B-106.

II. Roll Call

The following persons were present: Ed Cassidy, John Clocker, Steve Dwyer, Kimberly Ferguson, Tina Gheen, Ed Grossman, Mike Nibeck, Bob Reeves, Reynold Schweickhardt, Ralph Seep, Jeffrey Seifert, Arin Shapiro, Rob Sukol, Chuck Turner and Andrew Weber.

III. Approval of Minutes from Last Meeting

Bob Reeves e-mailed the minutes from the last meeting to the group. The minutes were approved as received, but an extension for edits was offered.

IV. Presentation on the Legislative Data Challenge

- Kimberly Ferguson, CRS, Tina Gheen and Andrew Weber, LOC, did a joint presentation on a Legislative Data Challenge proposal. The presentation focused mainly on the "big picture" items, and highlighted the Library's role in facilitating and administering the challenge. The broad goals, which are also aligned with the Library's goals are:
    - Cooperative exchange of legislative data
    - Leverage and extend existing standards
    - Create new metadata standards to handle exchange
    - Collaboration
    - Insight into the legislative process
    - Transparency
- The Library has carefully gone about vetting the legal and eligibility aspects of the data challenge.
- To administer the challenge, the Library plans to use Challenge.gov, which is a web platform sponsored by GSA. The website has been in use since 2008, and determined to be a great way to streamline the overall management of the data challenge. Many agencies have used the site to host similar contests. GSA recommends that the challenge is kept open for a minimum of 3 months.
- Bob Reeves e-mailed the group a link to the webpage used by the UK Parliament for a hackathon that they held late last month (Nov.).

- Although the Library has done a good job putting the challenge together, there is still some upfront work that must be done before moving forward. Some of the issue items discussed were:
    - Whether the international data standard can support our data standard
    - Expectations from the challenge
    - Use of DTDs vs. schemas
    - Building something with a backwardly compatible tool
    - Constellation prize for winner of the challenge
    - Next steps

V. Data Dashboard Ideas Discussion

- The bulk data group was not able to discuss data dashboard ideas as time ran out once again. Bob Reeves received feedback from GPO on website ideas. He asked the group to only submit their most popular sites and call categories for the sample web page. Suggestions should be sent to Toni Coverton. Discussions will resume during the next meeting as no presentations are scheduled.

VI. Project Updates

- Bob Reeves updated the group on projects currently underway:
    - The presentation on the new features that are part of the Phase 2 roll-out of docs.house.gov will take place sometime in January 2013.
    - The Bulk Data Task Force has to write a report to Appropriations Committee on its activities and findings. Bob Reeves hopes to get a draft submission to the group for review and feedback by end of next week. ***The report must be submitted to the committee by December 31.***
    - Plans are under consideration to invite the outside group back for a meeting in February 2013.
    - Coming Soon - - Floor Summaries will be available in bulk data XML format by Congress/Session in addition to the single day downloads soon.

VII. Schedule of Next Meeting

The next task force meeting will be held on December 20 at 11:00 AM in Cannon B-106. **We WILL NOT have a meeting this week!**

VIII. Adjournment

The meeting was adjourned at 12:02 PM.

Draft meeting notes taken by: Toni Coverton, Office of the Clerk

Legislative Branch Bulk Data Task Force

Draft Meeting Notes

Thursday, December 20, 2012 (11:00 AM - 12:00 PM)

I. <u>Call to Order</u>

Bob Reeves called to order the meeting of the Legislative Branch Bulk Data Task Force at 11:04AM on December 20, 2012 in Cannon B-106.

II. <u>Roll Call</u>

The following persons were present: Ed Grossman, Lyle Green, Selene Knoll, Ethan Lauer, Mike Nibeck, Jon Quandt, Bob Reeves, Reynold Schweickhardt, Ralph Seep, Jeffrey Seifert, Arin Shapiro, Andy Sherman and Rob Sukol.

III. <u>Approval of Minutes from Last Meeting</u>

Bob Reeves e-mailed the minutes from the last meeting to the group. The minutes were approved as received, but an extension for edits was offered.

IV. <u>Brief Presentation on the House Bills Bulk Data</u>

- Selene Knoll and Jon Quandt, GPO, did a brief presentation to update the group on the House Bills Bulk Data project. The presentation mainly focused on:
    - **Scope of Bulk Data**
        - *House Bill XML files are provided to GPO, the XML file contains information about the current chamber that is considering the legislation and the original chamber that introduced the legislation. This information is extracted from the bill XML files when they are processed by FDsys and is stored in metadata. The metadata is then used to automatically populate the Bulk Data based on configurable rules.*
        - *Legislative measures considered in the House may include Senate bills that have been referred in the House.*
    - **Congressional Bills (House)**
        - *The FDsys Bulk Data site will provide access to House Bill XML files beginning with the 113th Congress and moving forward.*
    - **Directories and User Guide**
        - *Each Congress directory will contain a session directory. The Bulk Data will broken down by Congress, by session and then by bill type. Each bill type directory will contain XML files along with a zip file. The directory and zip files are automatically updated as new bill files are made available on the main FDsys website, which generally happens twice a day.*

- Resources Directory
  - *A resources directory is also available, and contains all resources files currently found on the xml.house.gov website.*
  - *The House Bill XML User Guide will also be in this directory. The user guide was modeled after the Federal Register XML User Guide and contains an authentication statement. Bob Reeves will circulate a copy to the group.*
  - *There was some discussion about whether to add the US Code DTDs to the resources directory and other questions around what's inside the documents with "xsd" file extensions. Bob Reeves will follow up with LCS and get back to the group.*
- Discussions are taking place with Leadership on the best way to announce the roll-out of bill bulk data.

V. Data Dashboard Ideas Discussion

- After several weeks of running out of time during meetings, the bulk data group finally had an opportunity to continue the discussion of data dashboard ideas. There was a lot of discussion and questions raised on style, functionality, purpose, etc. Some of the items discussed were:
  - Who is the audience?
  - What are people looking for? What seems to be of interest?
  - What's the main goal/purpose of the site?
  - How do we package the information into a directory of available resources?
  - Should we combine searches and sites?
  - Do we prioritize links and change some of the wording?
  - Do we keep links/logos as a broad site, at high level?
  - How do we decide/govern which links should be included?
- Revisions will be made to the sample web page based on some of these suggestions and available for further discussion at a future meeting. Bob Reeves asked the group to continue to submit feedback to Toni Coverton.
- Follow-up items:
  - Each agency should review their website statistics to see where/what visitors are looking for.
  - It was also suggested that we review Data.gov as a possible data dashboard model.

VI. Project Updates

- Bob Reeves updated the group on projects currently underway:
  - Chuck Turner and Bob Reeves met with the LOC to ask them to move forward with the Bill Summary and Data Challenge initiatives. More analysis is required in both efforts.
  - A presentation on the new features that are part of the Phase 2 rollout of the Committee Project (docs.house.gov) will take place sometime in January 2013.

- We are considering another meeting with the outside group. The thought is to use the meeting to educate developers on the new bill text bulk data and Phase 2 of docs.house.gov.
- The official release of the new Historian website will be early January.
- Bob Reeves e-mailed a draft copy of the Activities Report to the group and asked for feedback by December 21. ***The report must be submitted to the Appropriations Committee by December 31. At some point, the committee may hold a hearing on it. The release of the Report will be up to the Appropriations Committee.***

VII. Schedule of Next Meeting

The next task force meeting is tentatively scheduled for January 10, 2013, but confirmation will be sent out in advance. This will be our first meeting in the New Year. Happy Holidays and see you in 2013!

VIII. Adjournment

The meeting was adjourned at 11:58 AM.

Draft meeting notes taken by: Toni Coverton, Office of the Clerk

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.

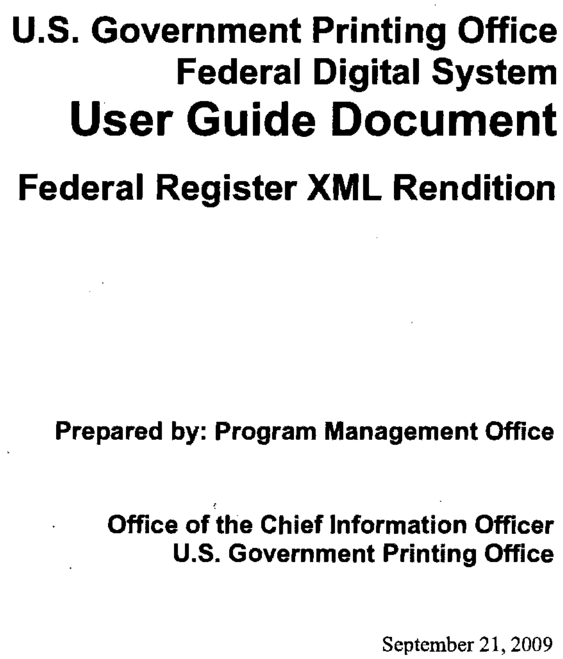
CRS also offers (and will continue to identify and analyze) the following policy matters for the Committee's consideration as it proceeds with determining next steps:

- ***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.



U.S. Government Printing Office
Federal Digital System
User Guide Document
Federal Register XML Rendition

Prepared by: Program Management Office

Office of the Chief Information Officer
U.S. Government Printing Office

September 21, 2009

Revision History

Revision	Date	Description
1.0	September 21, 2009	Version 1.0

Contact Information

For any questions regarding this document contact pmo@gpo.gov.

Table of Contents

1. Introduction

The U.S. Government Printing Office (GPO) and the National Archives' Office of the Federal Register (OFR) partnership is offering bulk data downloads of Federal Register files to the general public via Data.gov and FDsys. This effort began when the President challenged Federal agencies to create a more open and transparent government, promote accountability, and provide information to citizens about what their Government is doing (see 74 FR 4685, January 26, 2009 at http://www.gpo.gov/fdsys/pkg/FR-2009-01-26/pdf/E9-1777.pdf). The Public Printer's letter of March 9, 2009 pledged to provide trusted information in whatever form is required to meet the President's objectives http://www.gpo.gov/pdfs/news-media/letter_030909.pdf.

In addition, the Office of the Federal Register coordinates with the Office of Science Technology Policy to ensure that the OFR/GPO partnership meets customer expectations. To follow through on our commitment, we are expanding and accelerating the development of FDsys to provide XML-structured content as rendered output. This will give users access to masses of data to reconfigure and redistribute as they wish to meet the specialized needs of their constituencies.

1.1. Purpose

The purpose of this document is to provide an overview of Federal Register XML files and associated schema. The FDsys Bulk Data repository at http://www.gpo.gov/fdsys/bulkdata/ contains the Federal Register in XML from 2000 to the present. Please see FDsys at www.fdsys.gov for access to the Federal Register in PDF and HTML formats.

1.2. Legal Status & Authenticity of Federal Register files via Data.gov

Q. What is the data set available for the Federal Register in XML?

A. Federal Register files in XML have been converted and simplified from the SGML rendition used for the printed publication. The Federal Register is available in XML starting with the year 2000, when GPO began composing Federal Register material in SGML. OFR/GPO plan to convert the remaining electronic data set for the Federal Register (1994-1999) at a future date.

Q. Are bulk data downloads of XML files offered on FDsys via Data.gov part of the official version of the Federal Register?

A. No, the XML-structured files offered for bulk download are not part of the official on-line format of the *Federal Register.* While GPO's XML files are based on the original source data submitted by Federal agencies, OFR markup, and GPO typesetting and composition markup in SGML, only the PDF and Text versions of *Federal Register* content on GPO Access and FDsys have legal status as parts of the official online format of the *Federal Register.* Additional development will be required before OFR/GPO can specify that XML files are a part of the official online edition of the *Federal Register.*

Q. Do OFR/GPO plan to include XML files as part of the official format of the Federal Register?

A. Yes. The current set of XML-structured material (minus graphics) is not yet characterized as part of the official online *Federal Register* format because the underlying data used to create a viewable rendition of the material must be thoroughly scrutinized and tested. The XML-structured files are derived from SGML-tagged data and printing codes, which may produce anomalies in display. For example, complex tabular material in XML files may not display as correctly composed objects equivalent to the tables that appear in the Text and PDF files. Tabular material that displays in an ambiguous or distorted manner could affect the substantive meaning of regulations and other documents.

Our goal is to clean up the XML data and develop associated style sheets to the point that the XML rendition can be characterized as a display format of the official online edition. We are also developing the means to embed graphics in the files, so that an XML version may be deemed both official and complete. The OFR will issue a *Federal Register Bulletin* to alert users when the XML-structured files are ready to be included as part of the official online edition.

Q. What is the legal basis for making determinations about official status?

A. The Federal Register Act established the Administrative Committee of the Federal Register (Administrative Committee or ACFR) as the regulatory body with authority to determine the format(s) of the official serial publication known as the *Federal Register* (44 U.S.C. Ch. 15). Under 1 CFR 5.10, the official formats approved by the ACFR include a paper edition, a microfiche edition, and an online edition. OFR and GPO carry out ACFR regulations by developing appropriate means to display Federal Register material in the official formats.

Q. When did the Federal Register first appear online?

A. The official online edition dates to 1994 when Congress authorized an online edition in Public Law 103-40 (the Government Printing Office Electronic Information Enhancement Act of 1993; codified at 44 U.S.C. 4101).

Q. Are Federal Register XML bulk download files digitally signed?

No, XML files available for download are not digitally signed. They can be manipulated and enriched to operate in the various applications that users may devise. GPO is evaluating technology that could be used to digitally sign XML files for future official editions posted on FDsys. Adding signed non-PDF files to FDsys would be an enhancement for FDsys users, but would not be used to restrict or adversely affect the XML bulk data downloads available to our customers.

Q. What does the term "digitally signed" mean?

A. Currently, GPO uses digital signature technology on PDF documents to add a visible Seal of Authenticity (a graphic of an eagle) to authenticated and certified documents. The technology allows GPO to secure data integrity, and provide users with assurance that the content is unchanged since it was disseminated by GPO. A signed and certified document also displays a blue ribbon icon to the left of the Seal of Authenticity and in the Signatures tab within Adobe Acrobat or Reader. When users print a document that has been signed and certified by GPO, the Seal of Authenticity will automatically print on the document, but the blue ribbon will not print.

Q. How reliable is the metadata and the underlying tagging in bulk data files?

A. The document markup and metadata found within bulk data files are generally reliable and complete. However, there are variations in this underlying data due to inconsistencies in the composition and typesetting process. As a result, some data-mining applications and user aids may not produce 100 per cent accurate results.

Q. What is the legal status of Federal Register user aids?

A.: Federal Register user aids, including, finding aids, indexes, search tools, metadata associations, and tagging schemes are not part of the legal text of the *Federal Register*. These ancillary features help users explore and extract data, but the official legal text stands on its own. No person should form absolute legal conclusions based on search results, finding aids, metadata associations, extractions of data, and the like. Ultimately, only the official text of the *Federal Register* may be relied upon as evidence in a court of law.

Q. What is the authenticity of Federal Register bulk data files after they have been downloaded to another site?

A. We cannot vouch for the authenticity of data that is not under OFR/GPO control. OFR and GPO are providing free access to Federal Register data via XML for display in various applications and mash-ups outside the FDsys domain. The OFR/GPO partnership does not endorse third party applications, and does not evaluate how our original legal content is displayed on other sites. Consumers should form their own conclusions as to whether the downloaded data can be relied upon within an application or mash-up. An application may link to the official *Federal Register* on FDsys to provide users with additional assurance.

Q. Do OFR and GPO assert any control over downstream uses of bulk data?

A. In general, there are no restrictions on re-use of information in Federal Register documents because U.S. Government works are not subject to copyright. OFR and GPO do not restrict downstream uses of Federal Register data, except that independent providers should be aware that only the OFR and GPO are entitled to represent that they are the providers of the official versions of the *Federal Register* and related Federal Register publications.

Q. How can re-publishers indicate the source of Federal Register data?

A. Re-publishers of Federal Register data may cite FDsys and OFR/GPO as the source of their data, and they are free to characterize the quality of data as it appears on their site. But private sector re-publishers are prohibited from using the seal of the National Archives and Records Administration (NARA) or stylized Federal Register logos identified in NARA regulations (36 CFR part 1200) on their products because that would unlawfully misrepresent the legal status of the material, and could falsely identify private organizations as entities of the Federal Government.

2. Schema Description

The schema chosen to represent the Federal Register is a simplified version of the SGML schema that is used as part of the print production process, with some presentation and print specific tags removed or collapsed, and then converted to well-formed XML. This schema was chosen for the following reasons:

1. It is a complete and faithful representation of the Federal Register, which matches most closely to the author's original intent.
2. It describes the data using semantic tags in a way that is appropriate to the Federal Register Domain. For example, <RULE>, <NOTICE>, and <AGENCY> are all tags in this schema.
3. It fully describes the structure of the Federal Register, including the large structure (parts, articles, corrections, table of contents, etc.), the document structure (titles, paragraphs, sections, etc.), and semantic structure (CFR references, agency names, contact information, amendment text, etc.)

Since the schema is not an authoring schema and the SGML to XML conversion process maintains the order of the tags, this allows the XML schema to be more permissive than if it were used for checking authored content.

The schema being produced for this effort describes the data as it actually occurs from the OFR. Documents are not being cleaned up because they do not match the schema; instead, the schema was selectively relaxed. Such an approach maintains 100% fidelity to the original data, and eliminates any errors that might occur in schema interpretation or further data manipulation.

The following table lists the SGML tags that were removed or collapsed into a source attribute in the Federal Register XML.

Purposed Fields to be Removed	**Fields to be Removed or Collapsed**	**Description**
EDITOR	Removed	Used to indicate content editor
FNC	Removed	Used to generate a new column
FNEP	Removed	Used to generate a new even page
FNOP	Removed	Used to generate a new odd page.
FNP	Removed	Force new page.
Q	Removed	Inserts vertical spaces
NPAR	Collapse to P	Used to generate a new paragraph where the <P> tag would create a run in entry.
P	Collapse to P	Normally used for paragraph. A paragraph here has the first line indented.
P1	Collapse to P	Paragraph, indented one em on left.
P-1	Collapse to P	Paragraph, turnovers indented one extra em on left.

P1-3	Collapse to P	Paragraph, indented one em on left, turnovers indented three ems on left.
P2	Collapse to P	Paragraph, indented two ems on left.
P-2	Collapse to P	Paragraph, turnovers indented two ems on left. Same as FP1-2, which should not be used.
P2-4	Collapse to P	Paragraph, indented two ems on left, turnovers indented four ems on left.
P-3	Collapse to P	Paragraph, turnovers indented three ems on left.
P-DASH	Collapse to P	Paragraph, the last line of which fills with low-line dashes
OLNOTE1	Collapse to OLNOTE1	Sets footnote for first overlay note
OLNOTE2	Collapse to OLNOTE1	Sets footnote for second overlay note.
OLNOTE3	Collapse to OLNOTE1	Sets footnote for third overlay note.
OLNOTE4	Collapse to OLNOTE1	Sets footnote for fourth overlay note.
OLNOTE5	Collapse to OLNOTE1	Sets footnote for fifth overlay note.
OLNOTE6	Collapse to OLNOTE1	Sets footnote for sixth overlay note.
FP	Collapse to FP	Flush paragraph
FP1	Collapse to FP	Flush paragraph, all lines indented one em.
FP-1	Collapse to FP	Flush paragraph, turnovers indented one em
FP1-2	Collapse to FP	Paragraph, first line indented one em and turnovers indented two ems
FP2	Collapse to FP	Flush paragraph, all lines indented two ems
FP-2	Collapse to FP	Flush paragraph, turnovers indented two ems
FP2-2	Collapse to FP	Flush paragraph, all lines indented two ems.
FP2-3	Collapse to FP	Paragraph, first line indented two ems and turnovers indented three ems
FP3	Collapse to FP	Flush paragraph, all lines indented three ems.
FP-DASH	Collapse to FP	Flush line that fills with low-line dashes.
FRP	Collapse to FP	Flush right material, actually held in 1 em from right margin
FRP0	Collapse to FP	True flush right material
HD	Collapse to HD	First level head in the following sections.
HD1	Collapse to HD	First level head in the following sections.
HD2	Collapse to HD	Second level head in the following sections.
HD3	Collapse to HD	Third level head in the following sections.
HD4	Collapse to HD	Fourth level head in the following sections.
HD5	Collapse to HD	Fifth level head in the following sections.
HD6	Collapse to HD	Sixth level head in the following sections.

HD8	Collapse to HD	Lowest level head in the following sections.
HED	Collapse to HD	The first head in the following sections.
HED1	Collapse to HD	Special first level head in text of Section.
THED	Collapse to HD	Head that turns sideways on the page.
TSECT	Collapse to HD	Section head that turns sideways on the page.
FNC	Removed	Used to generate a new column

2.1. Sections Available in XML

The following are currently available in Federal Register XML:

- Contents
- Rules and Regulations
- Proposed Rules
- Notices
- Corrections

The following are not currently available in Federal Register XML:

- Front Matter (e.g. cover page)
- CFR Parts Affected
- Reader Aids
- CFR Checklist
- CFR Issuances
- Table of Effective Dates
- Graphics

2.2. Sections, Parts, and Articles

This section describes the top-level structure of a Federal Register XML file.

The XML schema, being a translated reproduction of the SGML schema, contains tags and content in the same order as they appear in the printed document. Major sections are grouped appropriately (<CNTNTS>, <RULES>, <PRORULES>, <NOTICES>, <NEWPART>, <CORRECT>), and all data and tags are represented – with the exception of the reduced tags from the previous section.

The XML tags and their descriptions of the schema above are shown below:

XML Tag	Description
FEDREG	The root tag for all document types in Federal Register publications. This tag includes children such as CNTNTS, RULES, PRORULES, NOTICES, NEWPART.

VOL	Contains the volume number for the publication.
NO	Contains the issue number of the volume.
UNITNAME	Contains the display name of the unit which follows, for example, "Notices", "Proposed Rules", "Rules", and "Presidential Documents".
CNTNTS	Contains table of contents pages for the Federal Register.
RULES	Used to start Rules section of the federal register. RULES may contain at least one <RULE>.
RULE	Used to start individual Rules in the Rules section.
PRORULES	Used to start Proposed Rules section of the Federal Register. Should contain at least one <PRORULE>.
PRORULE	Used to start individual Proposed Rule of the Federal Register.
NOTICES	Starts Notices section of the Federal Register. Should contain at least one <NOTICE>.
NOTICE	Start individual Notice within Notices section.

An abbreviated example of the overall section and structure is below:

```
<FEDREG>
  <VOL>65</VOL>
  <NO>21</NO>
  <UNITNAME>Contents</UNITNAME>
  <CNTNTS>
    ... THE CONTENTS OF THE TABLE OF CONTENTS ...
  </CNTNTS>
  <DATE>Tuesday, February 1, 2000 1-3-00</DATE>
  <UNITNAME>Rules and Regulations</UNITNAME>
  <RULES>
    <RULE> ... THE CONTENTS OF THE RULE ... </RULE>
    .
    .
    .
  </RULES>
  <UNITNAME>Proposed Rules</UNITNAME>
  <PRORULES>
    <PRORULE>  ... THE CONTENTS OF THE PROPOSED RULE ... </PRORULE>
    .
    .
    .
  </PRORULES>
  <UNITNAME>Notices</UNITNAME>
  <NOTICES>
    <NOTICE>  ... THE CONTENTS OF THE NOTICE ... </NOTICE>
    .
    .
    .
  </NOTICES>
  <NEWPART>
      ... THE CONTENTS OF THE PART ...
  </NEWPART>
  .
  .
  .
```

```
</FEDREG>
```

2.2.1.XPath Examples

The schema allows for a wide variety XPath commands for extracting items:

//RULE → Output all rules

(//PRTPAGE/@P)[0] → Get the first page number

(//PRTPAGE/@P)[last()] → Get the last page number

//RULE[descendant::PRTPAGE/@P = '12627] → Get rule which contains page 12627

//RULE[PREAMB/AGENCY = 'NUCLEAR REGULATORY COMMISSION']

→ Get all rules for the nuclear regulatory commission

2.3. NEW PART Section

The NEWPART section holds the contents of a part which often includes a part title and a list of notices, presidential documents, proposed rules, rules, and section names. A part number will always accompany the part title. In some cases, the specified rule will contain a preamble which can contain everything up to, but not including, the supplementary information.

The XML tags and their descriptions of the schema above are shown below:

XML Tag	Description
NEWPART	Used to start a new part of the Federal Register.
PART	Part group or container tag for search and retrieval purposes.
PARTNO	Sets the part number on title page of new part within volume.
PTITLE	Part title for new part within volume.

An abbreviated example of the NEWPART section and structure is below:

```
<FEDREG>
   .
   .
   .
  <NEWPART>
    <PTITLE>
      <PRTPAGE P="47239"/>
      <PARTNO>Part IV</PARTNO>
      <PRES>The President</PRES>
      <PNOTICE>Notice of July 28, 2000—Continuation of Emergency</PNOTICE>
    </PTITLE>
    <PRESDOCS>
      .
      .
      .
    </PRESDOCS>
  </NEWPART>
</FEDREG>
```

2.4. The Contents of Article

The contents of article tags are the ones that usually describe the majority of the data in the Federal Register. For example, XML tags such as HD and P are often used to start the head sentence in the following section while P tag is used to describe the rest of the text. The flush paragraph or FP is used to denote either large or small text depending on where it is being used. In addition, the majority of the sections are also following by Federal Register doc number as well as an agency billing code.

XML Tag	**Description**
HD	Used to start a new part of the Federal Register.
P	Part group or container tag for search and retrieval purposes.
DATES	Used to describe dates such as effective, applicable, and comment dates.
FP	Flush paragraph.
FRDOC	Federal Register doc number. It is used to track when entries were submitted for publication.
BILCOD	Agency billing code number.
EDNOTE	Used for editorial notes.

The contents will roughly have the same structure

```
<FEDREG>
   .
   .
   .
  <CNTNTS>
    ...
    <AGCY>
      <HD>Agricultural Marketing Service</HD>
      ...
      <SEE>
        ...
        <P>Farm Service Agency</P>
        ...
      </SEE>
      ...
    </AGCY>
    ...

 </CNTNTS>
   .
   .
   .
  <RULES>
    <RULE>
      .
      .
      .
      <EDNOTE>
        <FP>The Farm Service has...</FP>
      </EDNOTE>
      <DATE>December 4, 2000.</DATE>
      <FRDOC>[FR Doc. 00-31252]</FRDOC>
      <BILCOD>File 12-5-00; 8:45 am </BILCOD>
    </RULE>
  </RULES>
  .
  .     .
  .
</FEDREG>
```

2.5. PREAMBLE Tag

The PREAMBLE can hold contents that describe the current page of the document as well as agencies, sub agencies, actions, summaries, RIN numbers, and CFR citations. In general, a preamble tag can hold everything up to but not including the supplementary information. To view a complete list of valid fields, please see the XSD schema.

The XML tags and their descriptions of the schema above are shown below:

XML Tag	**Description**
PREAMB	Used to start the preamble. The preamble contains everything up to but not including the supplementary information.
PRTPAGE	Used to identify the page number in all documents in Federal Register
AGENCY	Used to identify the agency name.
CFR	Used to identify Code of Federal Regulation citation.
RIN	Used to identify regulatory information number.
SUBJECT	Used for title of subject.
ACT	Used to identify an action.
SUM	Used to identify a summary.
ADD	Used to describe contact addresses.
FURINF	Used to describe "FOR FURTHER INFORMATION CONTACT:".
SUPLINF	Used to describe supplementary information.
LSTSUB	Used to start a list of subject section.
REGTEXT	Used to designate regulatory information that will be inserted into the CFR.

An abbreviated example of the PREAMB section and structure is below:

```
<RULE>
  <PREAMB>
    <PRTPAGE P="12627"/>
    <AGENCY TYPE="F">NUCLEAR REGULATORY COMMISSION</AGENCY>
    <CFR>10 CFR Part 2</CFR>
    <RIN>RIN 3150-AI08</RIN>
    <SUBJECT>Interlocutory Review of Rulings on Requests...</SUBJECT>
    <ACT><HD SOURCE="HED">ACTION:</HD><P>Final rule.</P></ACT>
    <SUM>  ... the summary of the rule ... </SUM>
    <DATES>  .... Description of effective or applicable dates ... </DATES>
    <ADD>  ... contact addresses ... </ADD>
    <FURINF> ... where to go for further information about rule ... </FURINF>
  </PREAMB>
    .
    .
    .
  <FRDOC>[FR Doc. E8-4768 Filed 3-7-08; 8:45 am]</FRDOC>
  <BILCOD>BILLING CODE 7590-01-P</BILCOD>
</RULE>
```

2.6. SUPLINF Tag

The SUPLINF tag holds supplementary information that describes Federal Register documents. This information may include the page number, list of subjects or regulatory text material from the Rules section. The list of subjects will also include a list of CFR numbers cited in the appropriate rules.

The XML tags and their descriptions of the SUPLINF schema are shown below.

XML Tag	Description
SUPLINF	Used to describe supplementary information.
LSTSUB	Used to start a list of subject section.
REGTEXT	Used to designate regulatory information that will be inserted into the CFR.

An abbreviated example of the SUPLINF section and structure is below:

```
<RULE>
...
<SUPLINF>
    ... Supplementary information ...
    <LSTSUB>  ... List of CFR subjects appropriate to the rule ... </LSTSUB>
    <REGTEXT PART="2" TITLE="10">
        ... Specific changes to the CFR (Title 10, Part 2 in this example) ...
    </REGTEXT>
    .
    .
    .
  </SUPLINF>
...
</RULE>
```

2.7. Presidential Documents

The PRESDOCS tag must have one or more presidential documents which may include determinations, executive orders, memos, notices, or proclamations. A presidential notice, for example, will often have a title and presidential signature associated with it. It may also include the place of issuance which is usually "The White House." These items are always followed by

Federal Register doc number, the date filed, and the billing code.

The XML tags and their descriptions of the schema above are shown below:

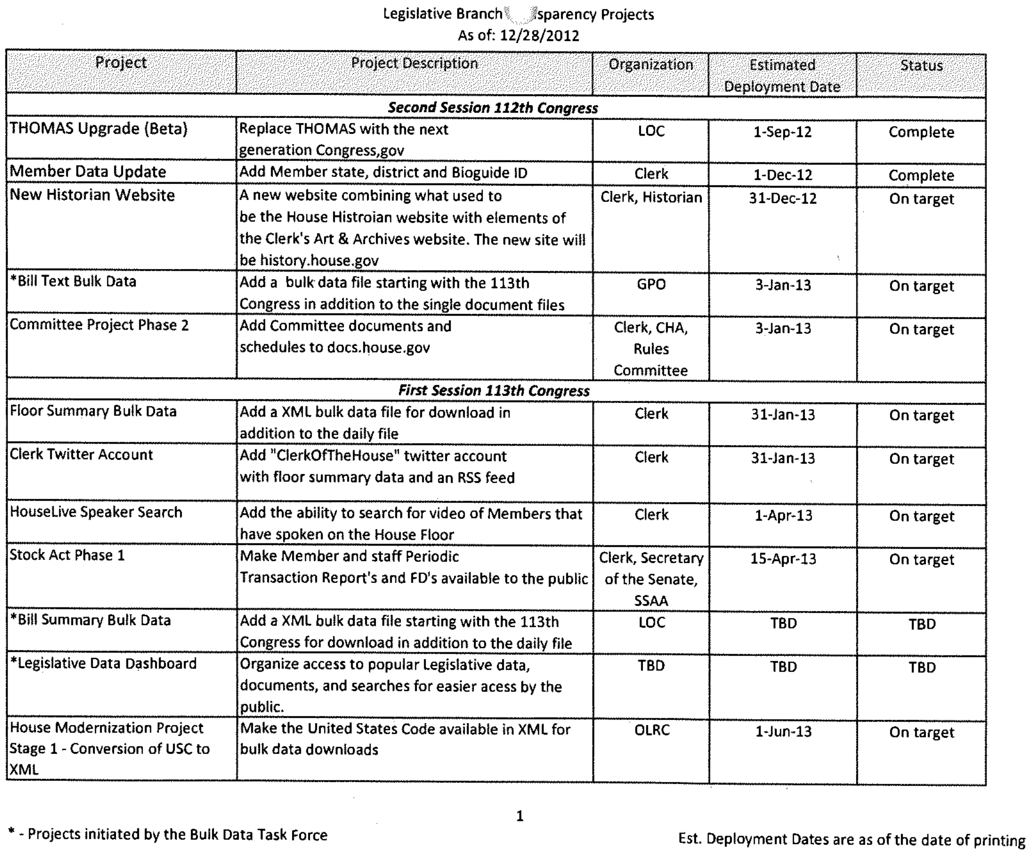
Node	Description
PRESDOCS	Used to start presidential documents section of the Federal Register.
PRESDOCU	Used to start individual Presidential document in the Federal Register.
PRESDOC	Used to start individual Presidential documents section of the Federal Register.
PRNOTICE	Notice of the Presidential documents.
DETERM	Presidential determination in Presidential documents.
EXECORD	Presidential Executive Order in Presidential documents.
PRMEMO	Presidential Memo in presidential documents.
PRORDER	Presidential Order in Presidential documents.
PNOTICE	Notice in Presidential documents.
TITLE3	CFR Title for Presidential documents.
PSIG	President associated with a Presidential document.
PLACE	Place of issuance for a Presidential document.
DATE	Date associated with the Presidential document.

An abbreviated example of the PRESDOCS section and structure is below:

```
<FEDREG>
...
  <NEWPART>
    <PTITLE>
      <PRTPAGE P="47239"/>
      <PARTNO>Part IV</PARTNO>
      <PRES>The President</PRES>
      <PNOTICE>Notice of July 28, 2000—Continuation… </PNOTICE>
    </PTITLE>
    <PRESDOCS>
      <PRESDOCU>
        <PRNOTICE>
          <TITLE3>Title 3—</TITLE3>
          <PRES>The President<PRTPAGE P="47241"/></PRES>
          <PNOTICE>Notice of July 28, 2000</PNOTICE>
          <HD SOURCE="HED">Continuation of Iraqi Emergency</HD>
          <FP>On August 2, 1990, by Executive Order 12722…</FP>
          <FP>This notice shall …<E T="04">Federal Register</E>…and…</FP>
          <PSIG>wj</PSIG>
          <PLACE>THE WHITE HOUSE,</PLACE>
          <DATE>July 28, 2000.</DATE>
          <FRDOC>[FR Doc. 00-19587</FRDOC>
          <FILED>Filed 7-31-00; 8:45 am]</FILED>
          <BILCOD>Billing code 3195-01-P</BILCOD>
        </PRNOTICE>
      </PRESDOCU>
    </PRESDOCS>
  </NEWPART>
…
</FEDREG>
```

3. Resources Directory

The resources directory in the Federal Register bulk data repository at http://www.gpo.gov/fdsys/bulkdata/FR/resources contains the current version of the XML schema, the XML stylesheet used to display the XML files in a browser on the FDsys website, and this user guide.



Legislative Branch [illegible]sparency Projects
As of: 12/28/2012

Project	Project Description	Organization	Estimated Deployment Date	Status
	Second Session 112th Congress			
THOMAS Upgrade (Beta)	Replace THOMAS with the next generation Congress,gov	LOC	1-Sep-12	Complete
Member Data Update	Add Member state, district and Bioguide ID	Clerk	1-Dec-12	Complete
New Historian Website	A new website combining what used to be the House Histroian website with elements of the Clerk's Art & Archives website. The new site will be history.house.gov	Clerk, Historian	31-Dec-12	On target
*Bill Text Bulk Data	Add a bulk data file starting with the 113th Congress in addition to the single document files	GPO	3-Jan-13	On target
Committee Project Phase 2	Add Committee documents and schedules to docs.house.gov	Clerk, CHA, Rules Committee	3-Jan-13	On target
	First Session 113th Congress			
Floor Summary Bulk Data	Add a XML bulk data file for download in addition to the daily file	Clerk	31-Jan-13	On target
Clerk Twitter Account	Add "ClerkOfTheHouse" twitter account with floor summary data and an RSS feed	Clerk	31-Jan-13	On target
HouseLive Speaker Search	Add the ability to search for video of Members that have spoken on the House Floor	Clerk	1-Apr-13	On target
Stock Act Phase 1	Make Member and staff Periodic Transaction Report's and FD's available to the public	Clerk, Secretary of the Senate, SSAA	15-Apr-13	On target
*Bill Summary Bulk Data	Add a XML bulk data file starting with the 113th Congress for download in addition to the daily file	LOC	TBD	TBD
*Legislative Data Dashboard	Organize access to popular Legislative data, documents, and searches for easier acess by the public.	TBD	TBD	TBD
House Modernization Project Stage 1 - Conversion of USC to XML	Make the United States Code available in XML for bulk data downloads	OLRC	1-Jun-13	On target

1

* - Projects initiated by the Bulk Data Task Force

Est. Deployment Dates are as of the date of printing

Legislative Branch Transparency Projects
As of: 12/28/2012

Project	Project Description	Organization	Estimated Deployment Date	Status
*Legislative Data Challenge	An international contest to convert US and British documents to Akoma Ntoso	LOC	TBD	TBD
Stock Act Phase 2	Automate the filing of FD's and PTR's Transaction Report's and FD's available to the public	Clerk, Secretary of the Senate, SSAA	1-Oct-13	On target
*Bill Text Bulk Data Phase 2	Add bulk data files for additional Congresses single document files	GPO	TBD	TBD
*Bill Summary Bulk Data Phase 2	Add bulk data files for additional Congresses single document files	LOC	TBD	TBD
Second Session 113th Congress				
MicroComp Replacement Project	A five year effort to eliminate microcomp	GPO	1-Oct-17	On target
House Modernization Project Stage 2 - Positive Law Codification System	Preparing and enacting a restatement of existing law	OLRC	TBD	TBD
House Modernization Project Stage 3 - Editorial Updating System	Replacing the functionality of a current system and produce XML output	OLRC	TBD	TBD

* - Projects initiated by the Bulk Data Task Force

Est. Deployment Dates are as of the date of printing

Reeves, Robert

From: Sherman, Andy
Sent: Friday, December 07, 2012 2:17 PM
To: Reeves, Robert
Subject: FW: Bulk data access to House legislative measures

Bob – is this the kind of thing you're looking for? Option 2 was the basis for our cost estimate for the project.

Andrew M. Sherman | *Chief Communications Officer* | Communications Office | ph 202.512.1991| mb 202.425.0123

GPO | OFFICIAL | DIGITAL | SECURE | 732 North Capitol Street, NW, Washington, DC 20401
Connect to us http://www.gpo.gov | http://www.facebook.com/USGPO | http://www.youtube.com/user/gpoprinter | http://twitter.com/#!/USGPO
Find Government information http://www.fdsys.gov | http://bookstore.gpo.gov | http://govbooktalk.gpo.gov

---

Option 1: Bulk Access via FDsys XML Sitemaps
Estimate: $0
Impacts:

- Implementation is already in place and being used by external bulk data users.
- The entire congressional bills collection on FDsys, including content files, metadata files, and authentication information, is currently available for bulk access via GPO's current implementation.

Approach:

- Utilize existing FDsys XML sitemaps to facilitate bulk access to House legislative measures using predictable URLs over HTTP.
  http://www.gpo.gov/smap/fdsys/sitemap_2012/2012_BILLS_sitemap.xml
  ```
  <url>
  <loc>http://www.gpo.gov/fdsys/pkg/BILLS-112hr4261ih/content-detail.html</loc>
  <lastmod>2012-03-31T05:45:12.000Z</lastmod>
  <changefreq>monthly</changefreq>
  <priority>1.0</priority>
  </url>
  ```
- Download XML content and metadata files using predictable URLs derived from the <loc> element within the FDsys XML sitemaps.
  http://www.gpo.gov/fdsys/pkg/BILLS-112hr4261ih/xml/BILLS-112hr4261ih.xml
  http://www.gpo.gov/fdsys/pkg/BILLS-112hr4261ih/mods.xml
  http://www.gpo.gov/fdsys/pkg/BILLS-112hr4261ih/premis.xml
- Parse the accompanying external FDsys MODS XML file for the <currentChamber> element to identify only House legislative measures.
  http://www.gpo.gov/fdsys/pkg/BILLS-112hr4271ih/mods.xml
  ```
  <currentChamber>HOUSE</currentChamber>
  ```
- Parse the accompanying external FDsys PREMIS XML file for the SHA-256 hash value contained in the <messageDigest> field for the XML file, and verify the hash value of the downloaded file against the hash value that is recorded in the FDsys PREMIS XML file.
  ```
  <objectCharacteristics>
  <fixity>
  <messageDigestAlgorithm>SHA-256</messageDigestAlgorithm>
  <messageDigest>587204c3269778027cfe19271a709167aafd87cd82a301c10697cf5b34b54ba8</messageDigest
  >
  <messageDigestOriginator>FDsys</messageDigestOriginator>
  </fixity>
  <format>
  <formatNote>Extensible Markup Language</formatNote>
  ```

```
</format>
</objectCharacteristics>
```

- Utilize the <lastmod> date within the FDsys XML sitemap file to determine recent changes to download files. <lastmod>2012-03-31T05:45:12.000Z</lastmod>

Option 2: Bulk Access via the FDsys Bulk Data Site
Estimate: $75,000
Impacts:

- Implementation is estimated to cost $75,000 and take three months to implement. An annual operating cost of $8,000 (+ or – 10%) would also be incurred.
- Implementation would only be available for House bills in XML from the 112th Congress forward.
- Implementation would require the use of additional storage because the XML bills will be stored on both FDsys and the FDsys Bulk Data site.

Approach:

- The FDsys Bulk Data site <http://www.gpo.gov/fdsys/bulkdata> will be configured for XML House bills using the directory structure /BILLS/112.
- FDsys Processing will be updated to read the value of the <congress> element and the <chamber> element in metadata for the BILLS collections on FDsys. Bills in XML format with the value of House and 112 will be added to the appropriate folder on the FDsys Bulk Data site.
- XML House bills along with associated files (e.g. DTD, style sheets, images) will be automatically added to this folder as a processing step within FDsys.
- The FDsys Bulk Data site will include a /BILLS/resources folder. The folder will contain an XML User Guide with authenticity information including an explanation of how to use the PREMIS XML files on FDsys to verify the authenticity of the XML bills on the FDsys Bulk Data site.

Andrew M. Sherman
Chief Communications Officer
U.S. Government Printing Office
732 N. Capitol Street, NW
Washington, DC 20401
202-512-1991

OFFICIAL DIGITAL SECURE

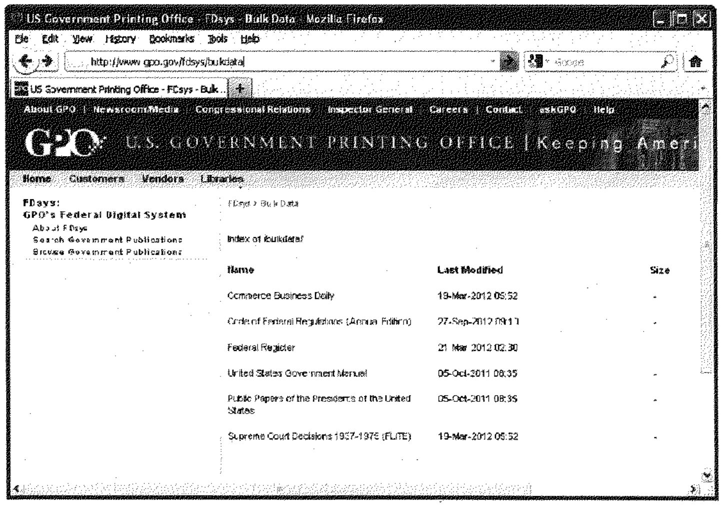
FDsys Bulk Data

Legislative Branch Bulk Data Task Force

Topics

- FDsys Bulk Data Site
- Federal Register XML Files
- House Bills Bulk Data
- Discussion

FDsys Bulk Data Site

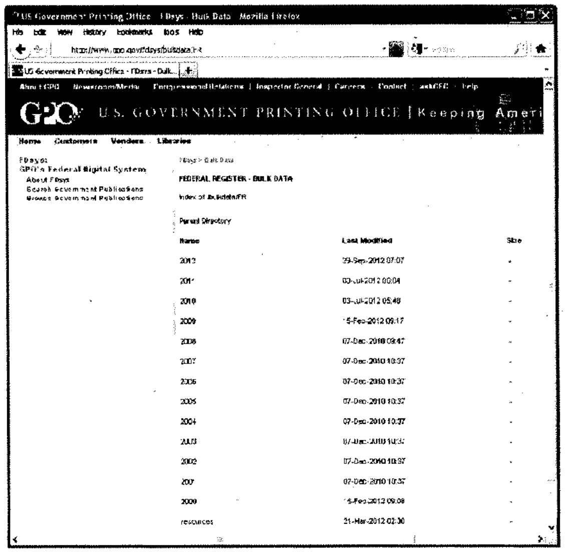

US Government Printing Office - FDsys - Bulk Data - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.gpo.gov/fdsys/bulkdata/
US Government Printing Office - FDsys - Bulk...
About GPO | Newsroom/Media | Congressional Relations | Inspector General | Careers | Contact | askGPO | Help
GPO U.S. GOVERNMENT PRINTING OFFICE | Keeping Ameri
Home Customers Vendors Libraries
FDsys:
GPO's Federal Digital System
About FDsys
Search Government Publications
Browse Government Publications
FDsys > Bulk Data

Index of /bulkdata/

Name	Last Modified	Size
Commerce Business Daily	19-Mar-2012 05:52	-
Code of Federal Regulations (Annual Edition)	27-Sep-2012 09:11	-
Federal Register	21-Mar-2012 02:30	
United States Government Manual	05-Oct-2011 08:35	-
Public Papers of the Presidents of the United States	05-Oct-2011 08:35	-
Supreme Court Decisions 1937-1975 (FLITE)	19-Mar-2012 05:52	-

- GPO currently provides bulk data download capabilities through the FDsys Bulk Data Site.
- The FDsys Bulk Data Site contains XML for select Office of the Federal Register (OFR) publications that are also currently available in PDF and text formats on the main FDsys website.

Federal Register XML Files



US Government Printing Office : FDsys - Bulk Data - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.gpo.gov/fdsys/bulkdata/FR
US Government Printing Office - FDsys - Bulk...
About GPO · Newsroom/Media · Congressional Relations | Inspector General | Careers · Contact · askGPO · Help
GPO U.S. GOVERNMENT PRINTING OFFICE | Keeping Ameri
Home Customers Vendors Libraries
FDsys:
GPO's Federal Digital System
About FDsys
Search Government Publications
Browse Government Publications
FDsys > Bulk Data

FEDERAL REGISTER - BULK DATA

Index of /bulkdata/FR

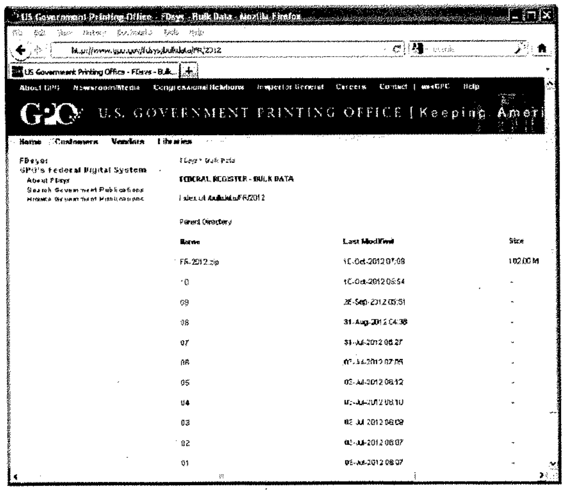
Parent Directory

Name	Last Modified	Size
2012	29-Sep-2012 07:07	-
2011	03-Jul-2012 00:04	-
2010	03-Jul-2012 05:48	-
2009	15-Feb-2012 09:17	-
2008	07-Dec-2010 09:47	-
2007	07-Dec-2010 10:37	-
2006	07-Dec-2010 10:37	-
2005	07-Dec-2010 10:37	-
2004	07-Dec-2010 10:37	-
2003	07-Dec-2010 10:37	-
2002	07-Dec-2010 10:37	-
2001	07-Dec-2010 10:37	-
2000	15-Feb-2012 09:08	-
resources	21-Mar-2012 02:30	-

- Federal Register XML files are available on the FDsys Bulk Data Site, and the files are grouped by year and then by month.
- The resources folder in the root directory contains a User Guide for the XML that includes a statement from OFR regarding the legal status and authenticity of Federal Register XML files.

Federal Register XML User Guide:
http://www.gpo.gov/fdsys/bulkdata/FR/resources/FDsys_OFR-XML_User-Guide-v1.pdf

Federal Register XML Files



US Government Printing Office - FDsys - Bulk Data - Mozilla Firefox

http://www.gpo.gov/fdsys/bulkdata/FR/2012

US Government Printing Office - FDsys - Bulk...
About GPO
Newsroom/Media
Congressional Relations
Inspector General
Careers
Contact | askGPO
Help
GPO U.S. GOVERNMENT PRINTING OFFICE | Keeping Ameri
Home
Customers
Vendors
Libraries
FDsys:
GPO's Federal Digital System
About FDsys
Search Government Publications
Browse Government Publications

FDsys > Bulk Data

FEDERAL REGISTER - BULK DATA

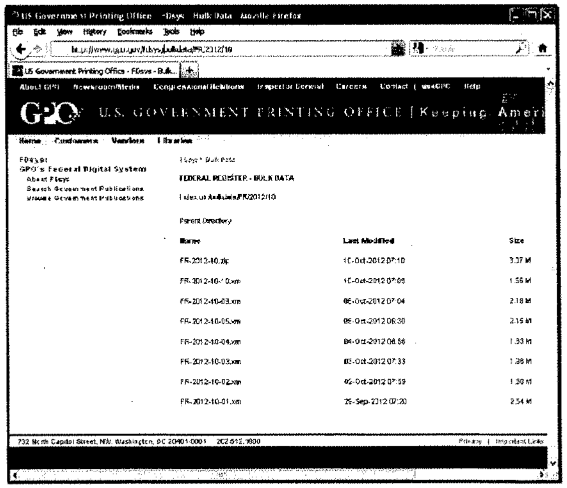
Index of /bulkdata/FR/2012

Parent Directory

Name	Last Modified	Size
FR-2012.zip	10-Oct-2012 07:09	102.00 M
10	10-Oct-2012 05:54	-
09	28-Sep-2012 05:51	-
08	31-Aug-2012 04:38	-
07	31-Jul-2012 06:27	-
06	07-Jul-2012 07:05	-
05	05-Jul-2012 06:12	-
04	05-Jul-2012 06:10	-
03	05-Jul-2012 06:09	
02	05-Jul-2012 06:07	-
01	05-Jul-2012 06:07	-

- All daily Federal Register XML files for an entire year are "Zipped-up" and made available on the FDsys Bulk Data Site.

Federal Register XML Files



US Government Printing Office - FDsys - Bulk Data - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.gpo.gov/fdsys/bulkdata/FR/2012/10
US Government Printing Office - FDsys - Bulk...
About GPO Newsroom/Media Congressional Relations Inspector General Careers Contact | askGPO Help
GPO U.S. GOVERNMENT PRINTING OFFICE | Keeping Ameri
Home Customers Vendors Libraries
FDsys.gov
GPO's Federal Digital System
About FDsys
Search Government Publications
Browse Government Publications
FDsys > Bulk Data

FEDERAL REGISTER - BULK DATA

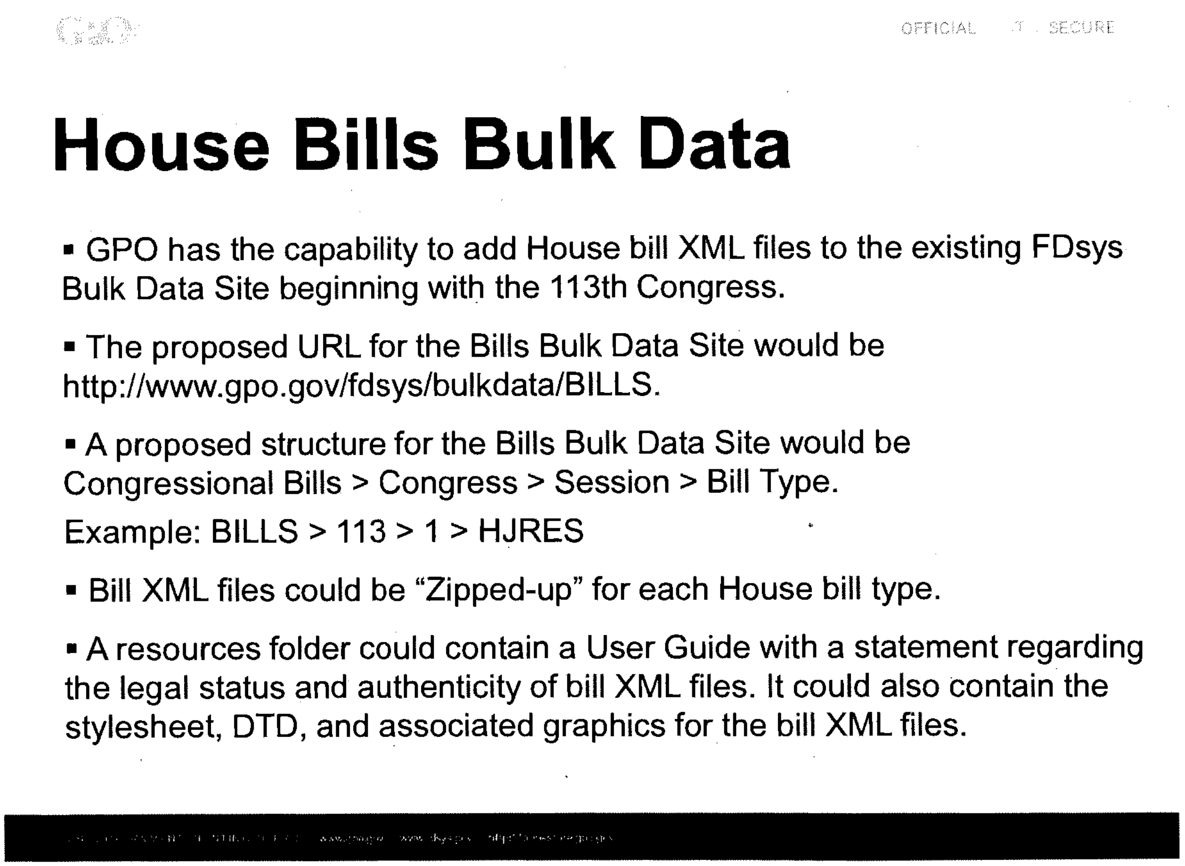
Index of /bulkdata/FR/2012/10

Parent Directory

Name	Last Modified	Size
FR-2012-10.zip	10-Oct-2012 07:10	3.07 M
FR-2012-10-10.xm	10-Oct-2012 07:09	1.56 M
FR-2012-10-09.xm	09-Oct-2012 07:04	2.18 M
FR-2012-10-05.xm	05-Oct-2012 06:30	2.15 M
FR-2012-10-04.xm	04-Oct-2012 06:56	1.33 M
FR-2012-10-03.xm	03-Oct-2012 07:33	1.38 M
FR-2012-10-02.xm	02-Oct-2012 07:59	1.30 M
FR-2012-10-01.xm	28-Sep-2012 07:20	2.54 M

732 North Capitol Street, NW, Washington, DC 20401-0001 202.512.1800

- All daily Federal Register XML files for an entire month are "Zipped-up" and made available in the month directory.
- Each individual daily Federal Register XML file is also available in the month directory.



OFFICIAL [illegible] SECURE

House Bills Bulk Data

- GPO has the capability to add House bill XML files to the existing FDsys Bulk Data Site beginning with the 113th Congress.

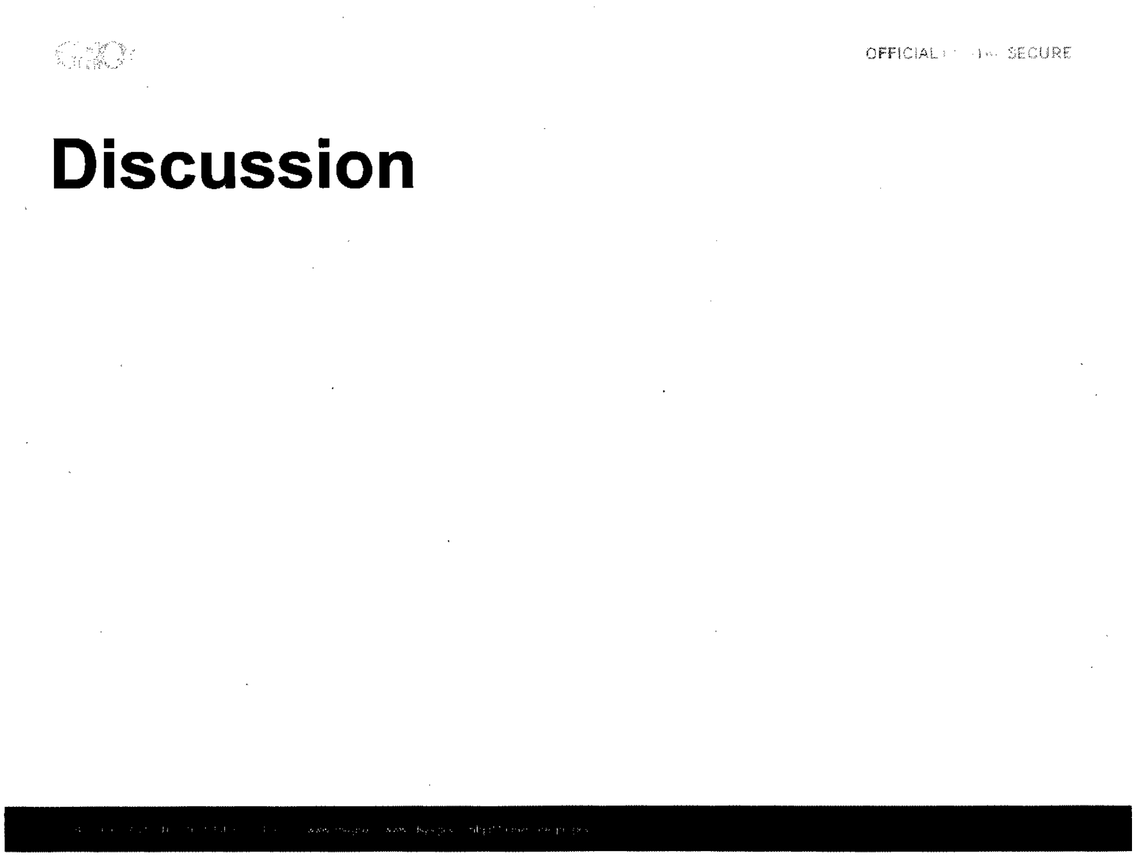
- The proposed URL for the Bills Bulk Data Site would be http://www.gpo.gov/fdsys/bulkdata/BILLS.

- A proposed structure for the Bills Bulk Data Site would be Congressional Bills > Congress > Session > Bill Type.

  Example: BILLS > 113 > 1 > HJRES

- Bill XML files could be “Zipped-up” for each House bill type.

- A resources folder could contain a User Guide with a statement regarding the legal status and authenticity of bill XML files. It could also contain the stylesheet, DTD, and associated graphics for the bill XML files.



OFFICIAL [illegible] SECURE

Discussion

The Library of Congress
Congress.gov
Bulk Data Access – Legislative Bill Summaries

Resources Estimate
October 25, 2012

Table of Contents

This document provides a rough order of magnitude estimate for the labor, hardware, and software required to establish a process within the Congress.gov system to generate bill summary data in XML format, create "bulk" data files that contain the XML data, and make the data files available to the public for download. A bill summary describes the most significant provisions of a bill text, and details the effects a bill may have on current law and federal programs, as defined in the legislative glossary on beta.congress.gov. Bill summaries are authored by CRS. The bulk data files will be generated on a daily basis and the archive of the bulk data will be arranged by Congress, session, and bill type (following how GPO is organizing bill data on FDSys)~~arranged by year, month and day~~.

There are two options for the delivery of the bulk data; this document outlines the costs of and considerations concerning both options. Estimates are provided for the implementation costs and for the ongoing yearly maintenance costs of each of the options. One option is to build a "portal" style page as part of the Congress.gov project that will allow anyone to access the files via a web browser. The second option is to provide the files only to GPO for public distribution through their FDSys bulk data repository. This solution would be built on the infrastructure used for the delivery of the Congress.gov system. The estimates assume any work would be managed through the full development life cycle that would include requirements analysis, system design, development, extensive testing, and a complete operations and maintenance plan.

Assumptions

- The bulk data extraction is an extension to the existing Congress.gov system.
- The bulk data will only contain bill summary information.
- Bill summaries are authored by CRS, so ownership and responsibility of the content of the summaries resides with the Library.
- The bulk data will NOT contain bill status, also called actions or status steps.
- The only summaries extracted for bulk download will be those for bills that originated in the House of Representatives.
- ~~The XML structure of the bulk data files (either the schema or the document type definition) must be approved by the Legislative XML Working Group.~~ The XML for the bill summaries and the resulting bulk data files shall be well-formed and must be approved by the Legislative Branch XML Working Group.
- The archiving of the content and delivery to GPO (if GPO is chosen as the delivery platform) will be performed by the existing media gateway product at the Library.

Discussion of Assumptions

Congress.gov has been built on a modern technical architecture that enables the development of new features and services. However, it was not designed specifically to facilitate the extraction of the data as XML documents for bulk download. It is possible that the continued development of Congress.gov that is planned in the upcoming years – which is focused on meeting the expert needs of Congress – will require the re-engineering of any bulk data extraction processes. To the greatest extent possible, these costs have been reflected in the estimates.

The estimates consider the effort for extracting, transforming and distributing only the summaries of bills that originated in the House of Representatives. Prior to initiating such a project, the Library would request written direction from the House to generate and release the information in bulk format. The Library would notify its Senate oversight committee of this activity.

If the scope of the data set to be extracted were to change, the estimates will need to be adjusted accordingly. Additionally, the Library would require a clear definition from the House that identifies the ownership of the data set and would also wish a concurrence with the Senate on that definition. The definition would inform the Library's view as to what authorizations would be appropriate prior to providing public access to the data in bulk format.

The full costs to the Library of supporting users of the bulk data is uncertain and cannot be accurately calculated for these estimates. Currently, third parties use "screen scraping" and other techniques to acquire data from THOMAS/beta.congress.gov. While the Library does not prevent such activities, it does not actively provide support for them, because they are outside the scope of providing a functional website for public use. If the Library purposely provides bulk download functionality, it anticipates that the third parties users will expect some level of support. Even if an "as is" type of disclaimer were provided, the Library foresees an increase in the number of inquiries and requests for assistance. The costs of maintaining documentation have been incorporated into the labor estimates; however, due to uncertainty in estimating the cost, the ongoing customer support activities for the new users – interested citizens, academics, interest groups, and information aggregators, and other businesses – have not been included.

If House bill summaries are released, the Library anticipates that demands for other information, such as short titles and relationships between the Congressional Record and bills to quickly follow. It is possible that some groups may try to leverage this action to drive demands for public dissemination of CRS reports, and perhaps other products as well. In addition to the support considerations for technical matters, the broader dissemination of certain types of products can create direct and indirect effects from unintended audiences. For example, CRS products are written solely for a congressional audience and are therefore tailored to the needs and context of Congress. If such products are purposely distributed to a much broader audience, they may be cited in more overtly political and less nuanced public discussion. In terms of direct effects, CRS could find itself fielding more inquiries from individual citizens, as well needing to clarify misrepresentations made by non-congressional actors. Such misrepresentations on controversial issues might spill-over to CRS work for Congress, requiring clarifications and repair of any reputational damage caused by others. In terms of indirect effects, over time such events can lead to subtle but substantive changes to the writing of reports and other products, as authors consider the potential reaction of outside audiences.

As noted in the Technical Requirements, below, the Library will provide a method for users to detect differences between the files downloaded in bulk and the files archived on the Library's servers. This method is expected to detect differences on a batch-by-batch, not "bill summary-by-bill summary", level. The Library notes that the legislative information, as a provided through a bulk extract, cannot be authenticated other than by comparison to the authoritative version maintained by the provider of the information. Once the information is hosted and "mashed up" by third parties, there exists no method for ensuring that the information has not been tampered with or innocently misinterpreted. Furthermore, distribution of bulk data will likely result in multiple alternative stores of legislative information that, to varying ~~degress~~ degrees are not as timely, and therefore as accurate, as Congress' primary systems. If there is an obligation to inform the general public to the risks of non-authoritative versions of the information, it has not been included in the estimates.

The XML structure of the bulk data files will remain consistent with the standards adopted by Legislative XML Working Group; changes to the XML structure may entail changes to the costs of the implementation and the maintenance.

The estimates address two options for delivery of the bulk data, one hosted by the Library and the other hosted by GPO. The Library's Office of the General Counsel (OGC) has raised a question as to whether bulk downloads for the public are consistent with our authority under 2 U.S.C. § 180 or whether it is instead provided for under GPO's authorizing statutes and appropriations. If the Library were to pursue the Library hosted delivery option, this question would have to be further explored.

High Level Process for Providing Bulk Data Access

- A data extraction routine will select appropriate bill summary data from the Congress.gov database.
- The extracted data will be transformed into the Legislative XML standard format.
- The extraction routine will notify a content management/delivery tool when new data is available.
- The delivery tool will create an archive copy of the content.
- The data will be made available either on GPO's site or on the Library's site.

Delivery Options

1. Library Hosted public access. The Bill Summary XML files will be located on a publicly accessible server. A portal will allow users to browse through the archive, navigating by year, month and day.
2. GPO Hosted. The Bill Summary XML files will be made accessible only to GPO for distribution through their existing bulk data infrastructure.

High Level Schedule

Analysis: 2 weeks
Development: 8 - 14 weeks
Testing: 2 weeks
Deployment: 1 week

Technical Requirements

- The XML files will conform to the existing Legislative Data XML standard.
- The solution will be built in such a way that the increased processing time will not materially impact the responsiveness of congress.gov to web users' requests.
- The solution will be implemented in such as way as it can be scaled to accommodate the required load (for the Library hosted option).
- The solution will utilize the existing media gateway infrastructure to manage the delivery of the files (for the GPO option).
- The ability to perform basic file integrity checking (via file level hashing) will be supported so that users can test whether that the data received in a download matches the data as stored on the Library's servers.

Estimated Resources

No hardware or software costs would be anticipated; labor costs would be as follows:

Initial Development and Deployment (Library hosted)

Project Manager, Analyst, Technical Architect/Expert, Software Engineer, Legislative domain expert, Testing	570 hours	$67,800.00

Yearly Ongoing Operations and Maintenance (Library Hosted)

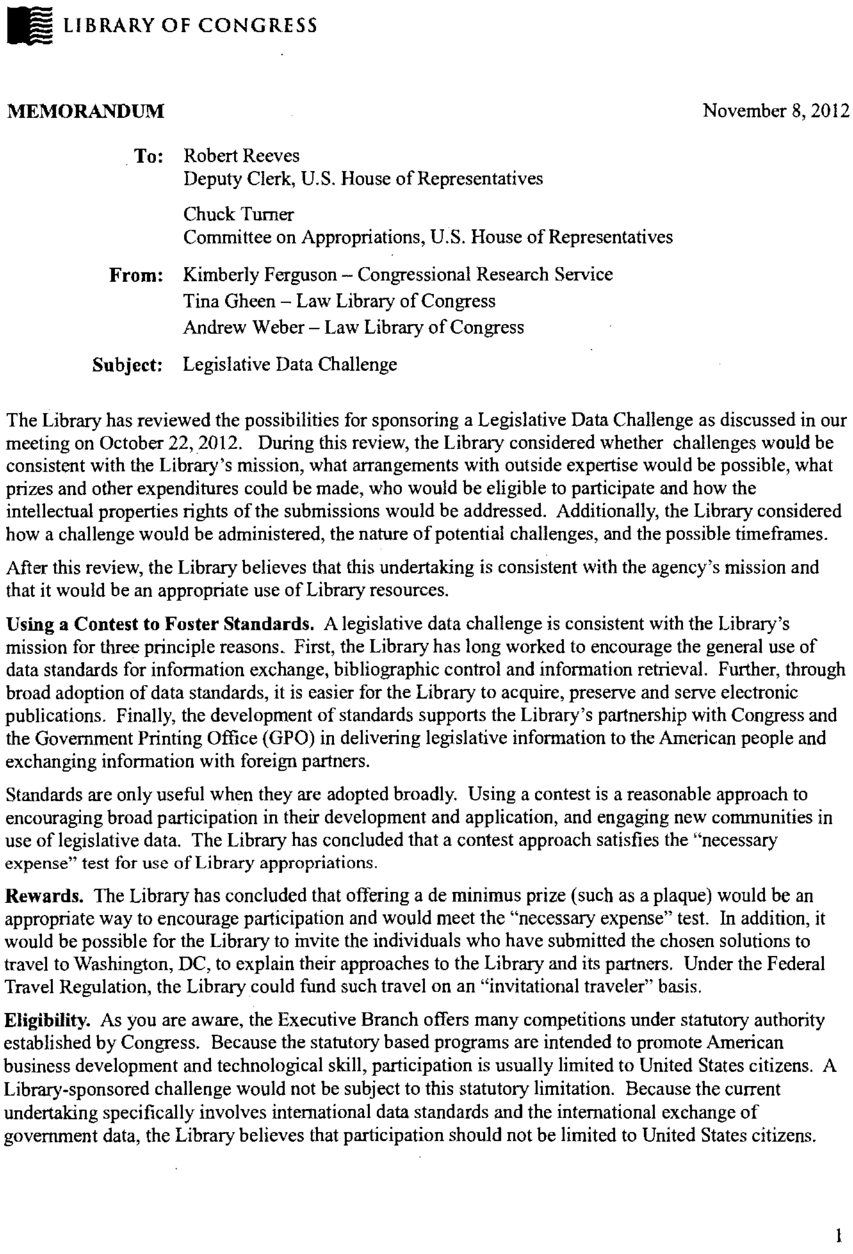
Project Manager, Analyst, Technical Architect/Expert, Software Engineer, Legislative domain expert, Testing	810 hours	$94,500.00

Initial Development and Deployment (GPO hosted)

Project Manager, Analyst, Technical Architect/Expert, Software Engineer, Legislative domain expert, Testing, Systems Engineering (GPO integration)	440 hours	$65,700.00

Yearly Ongoing Operations and Maintenance (GPO Hosted)

Project Manager, Analyst, Technical Architect/Expert, Software Engineer, Legislative domain expert, Testing	700 hours	$83,200.00



LIBRARY OF CONGRESS

MEMORANDUM November 8, 2012

To: Robert Reeves
Deputy Clerk, U.S. House of Representatives

Chuck Turner
Committee on Appropriations, U.S. House of Representatives

From: Kimberly Ferguson – Congressional Research Service
Tina Gheen – Law Library of Congress
Andrew Weber – Law Library of Congress

Subject: Legislative Data Challenge

The Library has reviewed the possibilities for sponsoring a Legislative Data Challenge as discussed in our meeting on October 22, 2012. During this review, the Library considered whether challenges would be consistent with the Library's mission, what arrangements with outside expertise would be possible, what prizes and other expenditures could be made, who would be eligible to participate and how the intellectual properties rights of the submissions would be addressed. Additionally, the Library considered how a challenge would be administered, the nature of potential challenges, and the possible timeframes.

After this review, the Library believes that this undertaking is consistent with the agency's mission and that it would be an appropriate use of Library resources.

Using a Contest to Foster Standards. A legislative data challenge is consistent with the Library's mission for three principle reasons. First, the Library has long worked to encourage the general use of data standards for information exchange, bibliographic control and information retrieval. Further, through broad adoption of data standards, it is easier for the Library to acquire, preserve and serve electronic publications. Finally, the development of standards supports the Library's partnership with Congress and the Government Printing Office (GPO) in delivering legislative information to the American people and exchanging information with foreign partners.

Standards are only useful when they are adopted broadly. Using a contest is a reasonable approach to encouraging broad participation in their development and application, and engaging new communities in use of legislative data. The Library has concluded that a contest approach satisfies the "necessary expense" test for use of Library appropriations.

Rewards. The Library has concluded that offering a de minimus prize (such as a plaque) would be an appropriate way to encourage participation and would meet the "necessary expense" test. In addition, it would be possible for the Library to invite the individuals who have submitted the chosen solutions to travel to Washington, DC, to explain their approaches to the Library and its partners. Under the Federal Travel Regulation, the Library could fund such travel on an "invitational traveler" basis.

Eligibility. As you are aware, the Executive Branch offers many competitions under statutory authority established by Congress. Because the statutory based programs are intended to promote American business development and technological skill, participation is usually limited to United States citizens. A Library-sponsored challenge would not be subject to this statutory limitation. Because the current undertaking specifically involves international data standards and the international exchange of government data, the Library believes that participation should not be limited to United States citizens.

In the Library's experience, there is interest and significant expertise in data standards and data conversion in the educational community, among non-profit organizations and in the commercial sector as well (such as in the publishing and information technology industries).

Intellectual Property. Because intellectual property is inherent in data standards and conversion software, the challenge documentation must be clear that the chosen solutions must be made available to the government and to the public either by being dedicated to the public domain or on an open source basis. (The Library recommends the Berkeley version of open source licensing, which basically only requires attribution.) The requirement to freely license IP rights would apply equally to all individuals, universities, businesses or other organizations who choose to participate in the challenge.

Judging. In order to bring appropriate expertise to bear, the Library would be able to enter into no-cost agreements with individuals to judge contest submissions and to identify to the Librarian those solutions that seem to be the most promising. The agreements with judges would need to include conflict of interest provisions in order to preserve the integrity of the process.

Proposed Challenges

The Library proposes to sponsor two legislative challenges:

Challenge 1: Map US and UK Bill Text Data Standards to Akoma Ntoso.

National, state, and local governments have established, or in the process of establishing, bill text markup standards through voluntary consensus standards development organizations. The map of United States, United Kingdom, and Akoma Ntoso standards will be a step toward defining a cross domain standard.

A panel of legislative data experts from the U.S. Congress, the U.K. Parliament, and the OASIS LegalDocumentML Technical Committee will evaluate submissions and select not more than one challenge winner. The challenge winner will demonstrate an application of voluntary consensus and de facto cross domain and domain specific standards, using the U.S., U.K., and Akoma Ntoso data models, as possible.

There are two objectives:

1. Map existing standards from voluntary consensus standards organizations (Legislative Branch XML Working Group, UK Parliament, Akoma Ntoso) for expressing cross domain metadata that is common to all open government data.

2. Propose new domain specific metadata as necessary.

Challenge 2: Map US Congressional Record and UK Hansards to Akoma Ntoso Standards

UK and Akoma Ntoso have developed markup standards for legislative debate through voluntary consensus standards. A map of UK and Akoma Ntoso standards, and proposed US standards will be a step toward defining a cross domain standard.

A panel of legislative data experts representing the U.S. Congress, U.K. Parliament, and the OASIS LegalDocumentML Technical Committee will evaluate submissions and select not more than one challenge winner. The challenge winner will demonstrate an application of voluntary consensus and de facto cross domain and domain specific standards, using the US, UK, and Akoma Ntoso data models, as possible.

There are two objectives:

1. Analyze which Akoma Ntoso and UK debate elements map to US data, and recommend new or adjusted elements

2. Analyze which Akoma Ntoso and UK elements map to non-debate content in the US Congressional Record, and recommend new or adjusted elements

Administering the Challenges. To administer these challenges, the Library anticipates using Challenge.gov, a web platform administered by the General Services Administration (GSA.) After reviewing the site and meeting with GSA representatives, the Library has concluded that the site will greatly simplify the process of receiving, processing and judging the challenge. Using this well-established channel will also introduce these legislative data challenges to a broad group of the public. The site is available without cost to all agencies that wish to sponsor contests, even those of us in the legislative branch.

The individuals who submitted the chosen solutions would be invited to Washington, DC, to explain their ideas at a meeting with the Library with other interested parties. While they are in Washington, we would hope to coordinate with your offices the presentation of a certificate or plaque to allow Congress to recognize their contributions.

Additionally, the Library would make the chosen solutions available on Congress.gov for the public.

Proposed Timeline. A proposed timeframe for the challenges is:

2012 Nov	Review details of both challenge tasks with 1) the Legislative Branch XML Working Group, 2) with OASIS LegalDocumentML Technical Committee. Kirsten Gullickson, Legislative Branch XML Working Group co-chair, has offered to help communicate with the OASIS technical committee, and 3) UK staff identified by John Pullinger.
2012 Dec	Finalize all challenge details.
2013 Jan	Announce and open challenge via Challenge.gov.
2013 April 30	Close challenge to further submissions.
2013 May	Judge challenge submissions.
2013 June	Announce results of challenge.
2013 TBD	Host meeting on standards, and in coordination with House, recognize the chosen submissions.

We are happy to discuss this with you at any time.

OPENING REMARKS—SERGEANT AT ARMS

Mr. ALEXANDER. Thank you. Mr. Irving.

Mr. IRVING. Good morning, Mr. Chairman, Ranking Member Wasserman Schultz, and members of the committee. I appreciate the opportunity to appear before you today to present the Sergeant at Arms request for fiscal year 2014. Before I begin, I would like to say that as Sergeant at Arms, it is indeed a unique privilege and honor for me to serve this institution, and I look forward to working with you and members of the committee. In the current fiscal environment, our office is acutely aware of the need to operate within tight fiscal boundaries. Our request has been crafted in the spirit of zero-based budgeting, where each division identified cost savings without jeopardizing the mission critical services provided to the House community.

113TH CONGRESSIONAL TRANSITION ACTIVITIES

My full testimony, which I have submitted for the record, contains my fiscal year 2014 budget request. In terms of ongoing efforts and initiatives, every division in the office of the Sergeant at Arms was recently involved in the transition to the 113th Congress. This includes the distribution of new Member identification pins and license plates, processing of approximately 14,000 congressional identification badges, and issuing over 7,000 parking permits to all authorized vehicles.

113TH CONGRESS HOUSE SECURITY ACTIVITIES

Furthermore, the employees of the Sergeant at Arms have supported, reviewed, and approved the security procedures for numerous special events, including the opening session of the 113th Congress, joint session of the Electoral College, the 57th Presidential Inauguration, and the annual State of the Union Address by the President. Support was also provided off-site to several issues retreats and the National Prayer Breakfast. Planning is currently underway for the annual Peace Officers Memorial Service and upcoming events on the west front lawn of the campus.

In closing, I would like to thank the committee again for their support and the privilege of appearing today. I assure you of my commitment and that of my entire office to provide the highest quality support for the House of Representatives, while maintaining the safest and securest environment possible. We will remain focused on security and preparedness, while maintaining the level of fiscal responsibility demanded by the House of Representatives. I will continue to keep the committee informed of my activities, and will be happy to answer any questions you may have. Thank you.

[The prepared statement of Mr. Irving follows:]

Statement of Paul D. Irving
Sergeant at Arms, U.S. House of Representatives
Before the
Subcommittee on Legislative Branch
Committee on Appropriations
Fiscal Year 2014 Budget Submission

Good morning Mr. Chairman, Ms. Wasserman Schultz, and members of the Committee. I appreciate the opportunity to appear before you today to present the Sergeant at Arms budget request for fiscal year 2014. Before I begin, I would like to say that as the Sergeant at Arms, it is indeed a unique privilege and honor to serve this institution. I look forward to working with you and the other members of this committee.

The Office of the Sergeant at Arms focuses its efforts on providing the maximum degree of support to Member offices coordinating security and protocol services as a highly integrated, flexible, and focused organization. The office is comprised of divisions that perform the duties mandated by the office: Police Services, Special Events and Protocol, Chamber Security, Parking Security, House Security, Information Services, and Emergency Management.

As Sergeant at Arms, I review and direct security matters relating to the House of Representatives, and as a member of the U.S. Capitol Police Board, I take part in establishing policies and guidelines to safeguard the Capitol complex, Members of Congress, staff, and the public conducting business and visiting the complex. I also serve as a member of the oversight board of the Office of Congressional Accessibility Services. This small, but important office is charged with providing and coordinating accessibility services for individuals with disabilities, including Members of Congress, officers and employees of the House of Representatives and Senate, and visitors to the Capitol complex.

There are a number of ongoing initiatives this office is involved with which I would like to bring to your attention today:

- Every division in the Office of the Sergeant at Arms has been actively involved in the transition to the 113th Congress. This includes the distribution of new Member and spouse identification pins and license plates, the processing of approximately 14,000 113th Congress ID badges, and issuing over 7,000 parking permits to all authorized staff vehicles.

- Employees of the Sergeant at Arms have supported, reviewed, and approved the security procedures for numerous special events, including the Opening Session of the 113th Congress, a Joint Session for the Counting of the Electoral College, the 57th Presidential Inauguration, the annual State of the Union Address by the President, and most recently, an unveiling Ceremony for the Rosa Parks Statue. Support was also provided off-site to several issues retreats and the National Prayer Breakfast. Planning is underway for the annual Peace Officers' Memorial Service and the upcoming concerts on the West Front lawn.

- In partnership with U.S. Capitol Police, we are continuing a strong and effective outreach program with Member offices regarding District Office security. We offer detailed guidance on best practices, providing information on how to obtain a thorough security review, coordination with local and state law enforcement for district security support, and how to coordinate security surveys when requested. With many new Member offices in the 113th Congress, I feel this process is even more critical. We will continue to provide this essential service to offices, while remaining cognizant of the need to provide cost effective recommendations as well as solutions to enhance the security of the Members.

- The Law Enforcement Coordinator Program (LEC) remains an important focus – it is something that should be an integral part of every District Office Security plan. Since the start of the 113th Congress, we have reached out to every Freshman Member office and continue to increase participation. As each of you know, LECs can and do provide an essential link to the local law enforcement community, enabling effective liaison and personal rapport with local and state law enforcement. The LEC program has clearly demonstrated its benefit to the institution, in a cost-effective manner.

- The Office Emergency Coordinator Program (OEC) continues to be a foundation in our overall emergency planning, and has been of great focus at the beginning of the new Congress. We continue to have a robust outreach program to all offices, offering emergency planning assistance, training, and guidance. OECs serve as the principal points of contact for their office in relaying emergency information and procedures. In the event of an incident, OECs are responsible for ensuring their office emergency procedures are carried out and assist with personnel accountability.

- We are in the process of developing and implementing on-line training for Law Enforcement Coordinators (LEC) and Office Emergency Coordinators (OEC).

- In the near future we will implement on-line Security Awareness "101" briefings which will be available to all House staff regarding foreign travel, operation security, and protection of personal identifiable information.

- Our division of House Security was established, in part, to become the repository for classified documents received by the House of Representatives. I am requesting funding to enhance the content management system used to store this material. Enhancements to this system will provide a vital backup ensuring robust data integrity, while reducing the chance of degradation or loss of data. We also look to increase the security of individual documents within the system through enhanced auditing, encryption, and authentication measures.

- We are preparing to implement a desktop "pop-up" notification tool for all PCs and a digital television display for all TVs on the Capitol Hill campus to be used for the quick dissemination of emergency messages.

- We continue to work on a comprehensive emergency management plan, in conjunction with the other Officers of the House of Representatives and Senate, and the U.S. Capitol Police.

In the current fiscal environment, every office is acutely aware of the need to operate within tight fiscal boundaries. Our request has been crafted in the spirit of zero-based budgeting where each division identified cost savings without jeopardizing the mission critical services provided to the House community.

In order to fund on-going efforts, the Office of the Sergeant at Arms is requesting $12,662,000 for fiscal year 2014. Of this amount we are requesting $9,091,000 for personnel expenses. The employees of the Office of the Sergeant at Arms are our most valued assets and have a shared responsibility in fulfilling the Sergeant at Arms mission. The performance of every Sergeant at Arms employee is vital to respond adequately to the needs of our stakeholders – the Members of Congress and the staff who serve them. While we are authorized for 132 FTE in fiscal year 2014, we are requesting funds for only 118 mission critical positions in Office of the Sergeant at Arms. We have reduced contractor support where possible and, as noted, are holding vacant positions open.

Non-personnel funding requested for fiscal year 2014 is $3,571,000. This funding will support travel, telecommunications, printing, other services, supplies and materials, and equipment.

Travel funding is primarily requested for the advance and support of official special events involving Members of Congress. Funding also supports some House emergency evacuation capabilities.

Telecommunications funding will support telephone, cell phone, air cards and wireless service for all divisions of the Sergeant at Arms.

Funding requested for Printing includes general printing needs as well as the preparation of emergency training materials.

Funding requested for Other Services is for contractual services in the areas of threat mitigation, force protection, counterterrorism, emergency preparedness, response and recovery, and the production of an on-line security refresher training.

Funding for Supplies and Materials is requested to purchase general office supplies, ID supplies, and miscellaneous supplies which include the necessary life-cycle replacement of

Parking Security uniforms, wireless devices, cellphones, and air-cards. This funding will also provide for the procurement of Member and spouse identification pins, license plates, and parking permits for the 114th Congress.

Funding for Equipment is requested to support hardware/software needs throughout all divisions, as well as required maintenance. Some highlights covered in this request include:

- Overdue lifecycle replacement of select PCs, laptops and other office equipment
- Purchase of critical equipment to support deployed House operations
- Annual maintenance and upgrades to several emergency planning software systems to be used in support of continuity of operations
- Purchase of certain equipment related to Continuity of Operations (COOP) activities, including portable badging/identification equipment and software

In closing, I would like to thank the Committee again for their support and for the privilege of appearing today. I assure you of my commitment – and that of my entire office – to provide the highest quality support for the House of Representatives while maintaining the safest and most secure environment possible. We will remain focused on security and preparedness, while maintaining the level of fiscal responsibility demanded by the House of Representatives.

I will continue to keep the Committee informed of my activities and will be happy to answer any questions you may have.

Thank you.