RC25240 (W1111-201) November 22, 2011
Computer Science

IBM Research Report

Many Bills: Engaging Citizens through Visualizations of Congressional Legislation

Yannick Assogba, Irene Ros, Joan DiMicco, Matt McKeon
IBM Research Division
One Rogers Street
Cambridge, MA 02142
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Many Bills: Engaging Citizens through Visualizations of Congressional Legislation

Yannick Assogba, Irene Ros, Joan DiMicco, Matt McKeon
IBM Research
One Rogers St, Cambridge, 02142
[yannick; iros; joan.dimicco]@us.ibm.com, matt@mattmckeon.com

ABSTRACT
US federal legislation is a common subject of discussion and advocacy on the web, inspired by the open government movement. While the contents of these bills are freely available for download, understanding them is a significant challenge to experts and average citizens alike due to their length, complex language, and obscure topics. To make these important documents more accessible to the general public, we present Many Bills (http://manybills.us): a web-based set of visualization tools that reveals the underlying semantics of a bill. Using machine learning techniques, we classify each bill's sections based on existing document-level categories. We then visualize the resulting topic substructure of these bills. These visualizations provide an overview-and-detail view of bills, enabling users to read individual sections of a bill and compare topic patterns across multiple bills. Through an overview of the site's user activity and interviews with active users, this paper highlights how Many Bills makes the tasks of reading bills, identifying outlier sections in bills, and understanding congressperson's legislative activity more manageable.

Author Keywords
Information visualization, text classification, government, legislation, government transparency.

ACM Classification Keywords
H5.3. Group and Organization Interfaces; Web-based interaction.

INTRODUCTION
The day after his inauguration in 2009, US President Barack Obama signed the Memorandum on Transparency and Open Government [1] stating that *transparency*, *participation*, and *collaboration* would be hallmarks of his administration. In the wake of this announcement, US government agencies began releasing data in digital formats for the general public to access and for application developers to build on top of. A number of non-profit organizations, such as the Sunlight Foundation[2], GovTrack.us[3] and MAPLight[4], emerged as clearinghouses for information and as advocates for a more data-driven approach to citizen and government communication.

One of the many issues these organizations seek to address is to help the average citizen navigate the complexity of US federal legislation. Knowing where to look in a bill for a particular issue or topic of interest is difficult because bills use specialized language and are, at times, quite long. A bill also may contain elements that are unrelated to its overall subject. One of our favorite examples of this appears in H.R. 627: The Credit Card Accountability Responsibility and Disclosure (CARD) Act of 2009, imposing transparency and disclosure requirements on credit card companies. The metadata provided by the Library of Congress categorizes this bill as a "Finance and Financial Sector" bill. However, embedded in Title V – Miscellaneous Provisions is a section titled Protecting Americans From Violent Crime that establishes the right of citizens to carry firearms in National Parks and Wildlife Refuges.

As this example shows, document-level metadata can hide important aspects of the underlying text. Motivated by this and a desire to make the content of legislative text less opaque, we have taken up the challenge of helping individuals read, explore and discuss US federal legislation. Our prototype, Many Bills, is a visualization and public website that enables anyone to search and explore the content of bills brought before the current Session of the Congress of the United States (running from Jan 3, 2009 - Jan 3, 2011). We use a machine learning approach to derive fine-grained topic overviews of bills and then visualize the results at various levels of detail including full-text browsing. Users can create custom views and embed them into a variety of other online media. Users can also explore bills based on which congresspeople sponsored them.

This paper presents Many Bills (screenshot in Figure 1), the techniques used to build and design the site, and initial responses from users. Through an overview of the site's user activity and interviews with active users, we highlight how Many Bills is used.

[1] *www.whitehouse.gov/the_press_office/Transparency_and_Open_Government/*
[2] *http://sunlightfoundation.com*
[3] *http://data.gov*
[4] *http://maplight.org*

Figure. 1. The homepage of Many Bills (http://manybills.us).

OUR DESIGN GOALS

Many Bills is part of a larger movement at the intersection of government and technology that aims to provide applications that transform government data into more approachable formats that citizens can understand, interpret, and communicate about. There are many examples of these types of "Gov 2.0" applications that have been built since Obama's memorandum on open government: the following are a small set of representative examples.

Two winners of the Sunlight Foundation's Design for America[5] challenge were Cool Kids at the White House[6], showing frequent White House visitors, and US Federal Contract Spending[7], showing federal spending versus media coverage. The visualizations are visually compelling and guide users to eye-opening conclusions based on the data.

Two other examples that allow users to interact even more with government data are Data Masher[8] and the National Obesity Comparison Tool[9]. These types of tools provide a limited collection of data, overlaid onto a geographic region, allowing users to filter and zoom into areas of interest.

One of our observations about the web applications built upon open government data is that the source data is often abstracted, providing a cursory summary of the data, allowing for only a few number of user interpretations. While engaging for the casual user, the applications do not cater to the curiosity of an advanced user-base who could potentially derive *new* meaning and insight from the data behind the visualization.

Our belief is that for citizens to become meaningfully engaged with government data, they need to be able to draw their own conclusions about it. Tools built around government data should provide guides and references to make finding the information they seek more apparent. So for our project, to support citizens in reading, understanding and interpreting legislation, we focused on these goals:

- Stay true to the text of the legislative bills while providing abstractions that make them more easily approachable: display a high-level view of a bill while providing access to its full text content.
- Create a visual interface that supports various levels of text abstraction to support different levels of interest in legislative text.
- Use standard web browser technologies to access a boarder audience, trading the rich capabilities of the desktop and plug-in based platforms for the wider reach of the web to citizens.

RELATED WORK

This research builds upon previous work on text visualization and other web applications that make US legislation available.

[5] *http://sunlightlabs.com/blog/2010/design-america-winners/*
[6] *http://www.nrftw.net/*
[7] *http://www.pitchinteractive.com/usbudget/*
[8] *http://www.datamasher.org/*
[9] *http://public.tableausoftware.com/views/contributorstoobesity/Eatyourvegtables*

Congressional Legislation Websites

There are three main websites that provide access to US federal legislation today. THOMAS [3], operated by the Library of Congress, is a comprehensive, web-accessible source of information on the activity taking place in Congress. Data includes bills & resolutions, public law, vote records, and legislator information. While bill text appears in full on THOMAS, it is presented in its raw form as a complete text document. No visual techniques are used to represent the data that might encourage informal engagement.

OpenCongress.org [2], a non-profit project run by the Participatory Politics Foundation, brings together legislative data, news media and individual blog posts, as well as public commenting on congressional activities. OpenCongress displays bills as formatted text without any visual abstraction or embellishment, similar to THOMAS. An additional enhancement is users' ability to comment on any portion of the bill.

A third site that provides open access to US legislative content is GovTrack.us [1], a tool by Civic Impulse, LLC. The site gathers congressional information from disparate official government websites (including THOMAS) and combines and cross-references the information to make it more reusable for third-party developers. GovTrack is a destination to read bills in a manner similar to that of THOMAS and OpenCongress. GovTrack's bill reading interface offers users a powerful embedding widget that can be used by bloggers and other media outlets to embed any portion of the bills within their site. GovTrack also serves as a legislative datasource to many applications, including Many Bills, because it provides a wealth of data to the developer community.

Connect2Congress [14] is an additional example of a visual analytics tool for exploring congressional activity, in this case voting patterns. It is a powerful and innovative effort to combine complex analytics and visualization methods to show the other side of legislative activity. However, without drill-down capability, users are not able to explore the data behind C2C's analysis.

Text Visualization

We considered many existing text visualization approaches for our challenge of visualizing legislative text. Our final approach differs from these examples of related work in a number of ways.

Visualizations such as Word Clouds [20] and Word Trees [21] can be useful for gaining insight into individual text documents. Word Clouds represent a document by relating features of individual words (such as frequency) while Word Trees explore the relationships of words to sentences. Both methods ignore the original relationship between the visualized units. In contrast Many Bills retains the structure of the original document, while augmenting it with metadata useful for topic based segmentation.

Wise et al.'s Themescapes [22] clusters documents based on thematic relationship in a 3-dimensional space. Havre et al.'s ThemeRiver [11] adds the dimension of time to show shifts in overall theme across documents. This class of visualization does not attempt to provide individual document views, which we see as critical to our goal in making legislative text more accessible to citizens.

Systems such as TileBars [12] FeatureLens [8], Digidock Explorer [4], and Jigsaw [18] support information retrieval tasks by providing a means to visualize the occurrence of terms or entities of interest across documents within a corpus. Many Bills also offers a higher level interpretation of the semantics of a particular section, as a guide to reading. While Literature Fingerprinting by Keim and Oelke [13] allows for visual comparison of multiple documents, the metrics displayed are not semantic in nature but rather are lower level linguistic features of text that are best suited to an expert analyzing a document corpus. Document Cards [19] and SmartNails [5] generate semantic, fixed-size thumbnails of documents. Each of these methods provide high-level summary information, but not a high level guide to the original contents of the documents.

Most related to Many Bills are Seesoft [9], ViewTool [6] and Compus [10] that each integrate full-text browsing into the visualization itself. Seesoft and Compus display a reduced view of documents as adjacent columns; coloring portions of the text according to various line-level metrics (SeeSoft) and human-annotated features (Compus). While similar in layout, Many Bills visualizes machine-generated semantic descriptors of document substructure, giving a higher level overview of documents. ViewTool combines a high level, topic-based single-document overview and a full-text view into a single 3-pane interface while Many Bills offers a view of multiple documents allowing users to compare topic distributions.

Plaisant et al.'s visualization of Emily Dickenson's correspondence [17] uses a human-assisted machine classification system to assign documents into one of two categories for further analysis. However, its heavy reliance on experts for both classification and interpretation, as well as the low granularity of its categories, limit its application to the more general problem we are attempting to address.

One example of a project visualizing government textual data is Parallel Tag Clouds, by Collins, et al. [7], explored a large government dataset of complex legal documents focusing on district court rulings. Using a combination of parallel coordinates and tag clouds, users can explore the common words appearing in each district. While this visualization offers insight into the prominent topics in each district, it does not allow advanced users to delve deeper into the dataset, a common trend among text visualizations.

IMPLEMENTATION

Many Bills is a web-based application built on the Ruby on Rails platform using HTML, CSS and Javascript on the front end. In this section we discuss the text analysis technique used and the visualizations available to users on the site.

Text Analysis

While legislation is accompanied by descriptive metadata, the only information provided about the different portions of a bill is within a table of contents. Yet summary information about the different portions of a bill could be useful for providing guidance to citizens on whether they want to read a bill, or which portions of a bill would be most interesting to them. Thus, in analyzing congressional legislation, we sought to discover the distribution of topics that a bill contains.

Congressional bills are structured hierarchically, and although the hierarchy varies across bills, they consistently contain a discrete unit, known as a *section*, that typically covers a single provision of legislation.

Our approach is to use a trained machine learning classifier to estimate the probability of a section of a bill being about a particular subject. To train the classifier we took advantage of the fact that each bill is assigned a *top subject* by the Congressional Research Service (CRS) of the Library of Congress. We use these subject assignments on a collection of over 100,000 bills from 2000-2009 to train a classifier that we then use to label individual sections. Our assumption is that the model that predicts the top subject of an entire bill can be used to predict the subject of an individual section. As new bills are proposed in this session of Congress, we run the classifier over the new bills' sections. To date, there are over 9000 unique bills on the website.

We use a maximum entropy classification algorithm provided by the MALLET toolkit [15] with an 80/20 training/test split and 10-fold cross-validation. No special parameters were used in tuning the algorithm, nor were bills pre-processed prior to classifier training. Our initial results are promising. For example, in the aforementioned Credit CARD Act of 2009, most sections carry the label "Finance and Financial Section," yet the section on gun rights is classified as "Public Lands." Each section receives a probability for each possible top subject provided by the CRS (approx. 80 different subjects). We do not attempt to classify sections with less than 50 words, as we found that the in those cases the results are rarely accurate.

We performed an initial evaluation of the quality of the predictions made by the classifier. Given the large number of sections that we have labelled (73,020 sections at the time of this evaluation), it is difficult to hand-rate a representative sample of sections. As a baseline measurement, we compared the overlap between the subjects we assign to the full set of subjects assigned to the bill by the CRS. In Table 1, we report at different levels of classifier confidence how well our classifications overlap with the CRS classifications for the bill. To calculate this in a manner that is sensitive to the different lengths of bills, we look at the subject we assign to each section in a bill and assign a point if that subject is in the set of subjects indicated by the CRS for that bill. We then divide the total number of points by the number of classified sections in that bill. This measures how many of our classifications match those that the CRS suggest should be present within the bill.

This analysis method does not take into account situations where we may do better than the CRS at suggesting a topic -- we are in fact penalized when this occurs. For example, in the Credit CARD Act, the subjects assigned to the bill are: Finance and Financial Sector, Administrative Law and Regulatory Procedures, Banking and Financial Institutions Regulation, Consumer Credit, Federal Reserve System, Government Information and Archives, and Interest, Dividends, Interest Rates. None of the subjects suggest content related to second amendment rights, national parks, or education on financial literacy: all topics that we are able to identify with our technique.

Admittedly the performance of the classifier as reported in Table 1 is not ideal, however this is an area we continue to work on improving. We hope to increase this performance by better modeling of the language used in the entire dataset and improving feature selection for the classifier. Future work will also gather classifications from our users. One might ask, given there is no ground truth for section classifications, why don't we use unsupervised techniques such as clustering. The primary reason we chose a supervised technique is the ability to present semantically interpretable labels to the user. With clustering, the output would inform us which sections differ from others, but wouldn't tell us why. In early experiments this turned out to be unsatisfactory from a comprehension perspective.

Table 1. Classification Results

Confidence Threshold	Classified Sections (Out of 73,020)	Section classifications that match CRS classifications.
> 0	100%	51%
> 0.20	92%	53%
> 0.40	82%	54%
> 0.60	69%	55%
> 0.80	52%	59%
> 0.99	22%	72%

Visualization Design

The Many Bills website provides two perspectives from which to explore legislation. The first, a document-centric view reveals the topical substructure of bills we have classified. The second provides a view of the different topics that legislators align themselves with and allows users to explore bills from a people-centric perspective.

Document-Centric Visualization

The analysis described previously generates a set of congressional bills where each section has been assigned one or more subjects, along with their associated probabilities. We assign each section a color based on its subject. As we have over 100 different possible subjects, it is impossible to create a color scheme that allows the user to visually distinguish subjects from one another. We

manually grouped semantically-related subjects together and assigned each group a color; for example, the subjects "Consumers," "Trade," and "Taxation" all fall under the "Economics" group, which is assigned a green color.

We render each bill's sections as vertically stacked blocks; a block's color corresponds to the top subject (the subject with the highest probability) assigned to that section. The height of a block is mapped to the length of that section in the bill. Inside each section's block, its title is displayed to provide a quick overview of the section's content. When a section is opened by clicking on it, or when a user mouses over it, the classifier's confidence score is displayed. The actual subject assigned to the section is shown by a small badge displayed to the left of the block; each badge is formed from the first few letters of its subject. Figure 2 shows a single bill as represented in Many Bills.

Figure 3 shows four bills in a user-generated collection about the extension of unemployment benefits. The document-centric visualization screen can display up to 50 bills on a single screen at a time, with larger collections spanning over multiple pages. Users can save any group of bills into an annotated collection, with a title and description, by adding them to a temporary storage area known as the 'bill tray'. These collections are browsable from the home page and can be shared with others.

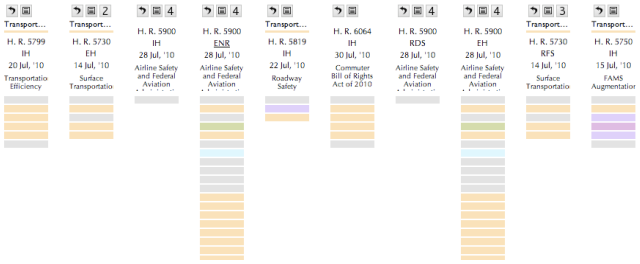

Figure. 4. Ten bills represented in "minified" mode where each section is a color rectangles.

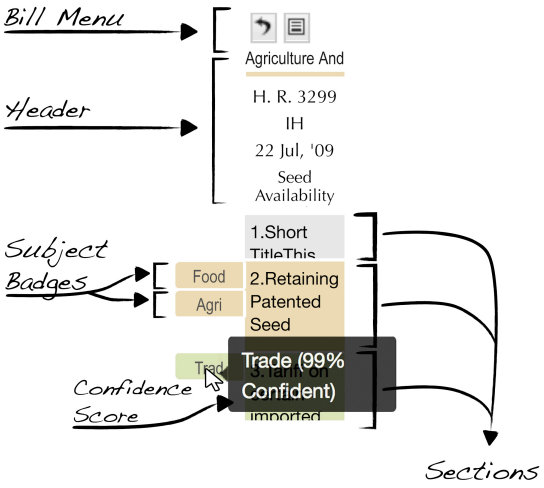

Figure. 2. How a single bill appears in Many Bills, with each section color-coded according to classification, marked with a classification badge.

The visualization supports several interaction methods. When a user clicks on a bill section, it expands and displays the full text of the section. This is in line with our primary goal of supporting users in accessing the original data. Furthermore, the entire contents of a bill can be read by clicking a button in the toolbar displayed above each bill. A color legend is provided for each collection page that displays all subjects visible on that page. Users can toggle specific colors on/off, selectively coloring or graying out sections associated with that subject. This enables users to quickly locate topics of interest in the current collection.

Users can switch the visible bills into "minified" mode, shown in Figure 4. In this view, each section is reduced to a rectangle with a fixed height, enabling users to compare patterns of subjects across bills of greatly varying length. Due to the extreme length of some of the more significant bills, this view is particularly valuable for compressing the visual length of a bill. In this mode users are encouraged to navigate the collection of bills through color and tooltips, which show each section's title and classification. Just as in the default view, a single click on a section expands the section to the full text.

Many Bills supports comprehensive search that includes subject, author, date, and full text. The search results appear in the visualization as in the same manner as any other collection, with sections containing search hits highlighted with a yellow bar.

People-Centric Visualization

An alternative visualization provided on Many Bills looks at legislation from the perspective of the people directly involved in the process of its creation and debate. For each Senator and House Representative we provide a page detailing their legislative activity from the current session.

The page comprises of three parts, first (Figure 5a) is a plot that displays every bill sponsored or cosponsored by a congressperson as a square, color coded by the CRS provided subject for that bill. Mousing over each square provides additional information about the bill and clicking on the

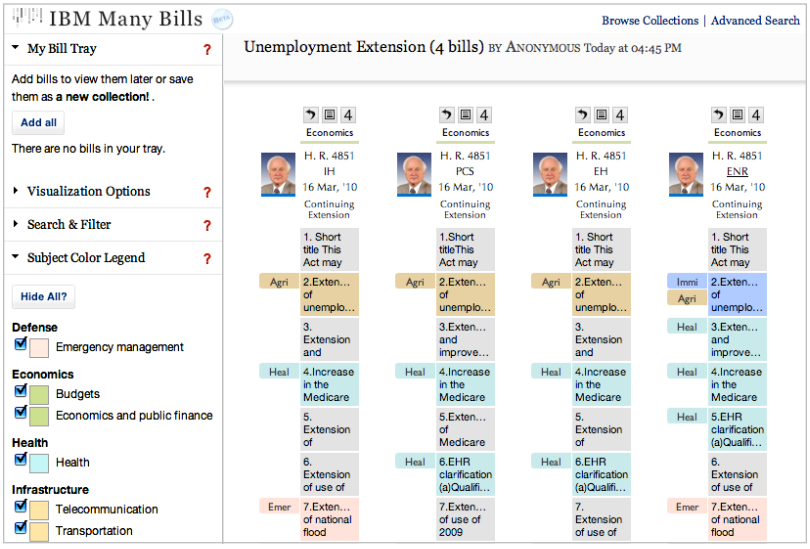

Figure. 3. Four bills in a user-generated collection about the extension of unemployment benefits.

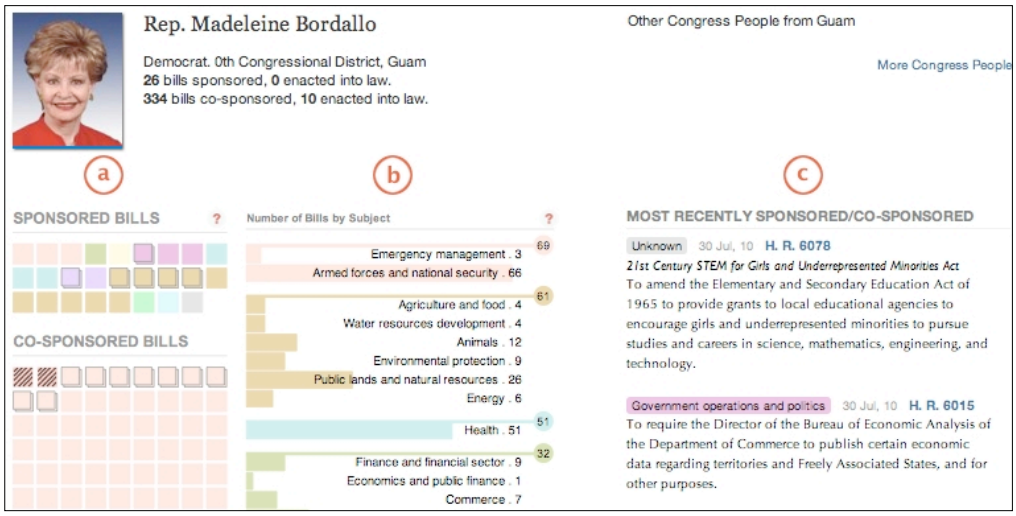

Figure. 5. A Congressperson page. a. Bill plot. b. Bar graph. c. Recent bills.

square takes you to the content of that bill. This plot allows one to get a summary of the various topics a congressperson is particularly active with. The squares are also shaded to indicate how far along in the process a bill is: hashmarks for passed, stacked squares for on the legislative calendar, and plain for introduced.

We also provide a bar chart (Figure 5b) that shows the exact proportions of bills sponsored or cosponsored per member. By employing the interaction techniques of brushing and linking, we allow users to highlight particular parts of the bar chart to filter the first plot of all the bills and highlight those in a particular subject of interest making them easier to pivot to.

The last major element on this page, is a list of recently sponsored or cosponsored bills for the congress person in question (Figure 5c). This list displays the titles of recently sponsored bills as well as their subject assignments from the CRS.

ILLUSTRATIVE EXAMPLES

The primary goal of Many Bills is to offer visual summaries of legislation that offer guidance on where to find interesting or unusual sections of bills. Several examples have provided validation of this and we believe show the various ways in which the tool can help citizens read bills in a more manageable way.

As discussed in our analysis of the classifier's performance, the section on firearms in the Credit CARD Act is correctly labeled as not relating to the "Finance and Financial Sector," but rather "Public Lands." Figure 6 shows how the bill appears within the document-centric visualization. The 'Protecting Americans from Violent Crime' section clearly stands out as a brown rectangle in a sea of green 'Finance' sections.

Another example is the controversial health bill that was signed into law in late 2009. This bill, H.R. 4872 Health Care and Education Affordability Reconciliation Act of 2010, makes substantial changes to health care regulation in the US and its length, of over 2000 pages, was mentioned in the media discussion. Viewing the bill in Many Bills highlights several aspects. As was often mentioned in the media, this bill contains substantial portions of education legislation and this is easily observed by turning off all colors except for those related to education. Instantly, a portion at the end of the bill is lit up, revealing the exact location of these education sections. We believe these dense documents can be made more digestible overall by providing such visual segmentation of a bill into constituent topical clusters.

Another aspect of this bill that dominated media and political discussion was its characterization as "government-run healthcare:" opposing politicians claimed the bill authorized government rationing of healthcare, referred to as "death panels [16]." In our visualization a deep pink color highlights sections relating to "Government and Politics." When exploring the bill, a pink section stands out from the blue "Health" color; section 141, which outlines the formation of a Health Choices Administration and a Health Choices Commissioner. Subsequent sections continue to outline the duties of this new part of the executive branch. It is our hope that making these key topics easier to find will aid citizens in discussing and reasoning about the issues based around the content of the bill, rather than pundits' politicized interpretations of the bill.

Another bill of interest is S. 22 ES, the Omnibus Public Land Management Act of 2009. This bill designates certain lands as components of the National Wilderness Preservation System, in addition to enacting other legislation related to public lands. It is also an extremely long bill with 389 sections. By turning off all of the categorization colors, and then turning on individual category colors, one can discover there are three sections labelled "Medicine." By expanding these sections to their full text, one finds the Christopher and Dana Reeve Paralysis Act, which funds a paralysis rehabilitation center. Just as the firearms provision was added to the Credit CARD Act, this provision deviates from the main bill topic, but may have been necessary to ensure passage of the bill.

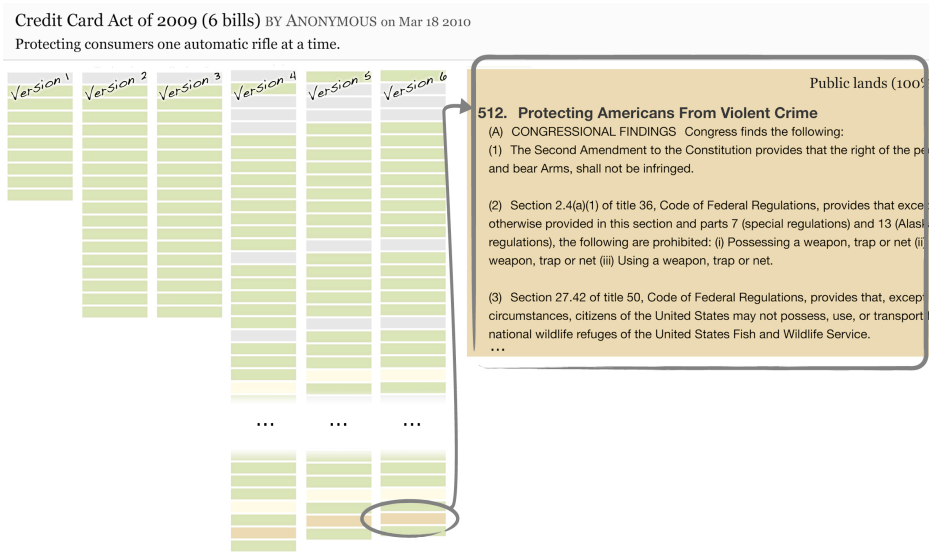

Figure 6. The Credit CARD Act of 2009, as it appears in Many Bills. The bill protects consumers from predatory lending practices by credit card companies. Additionally, it allows citizens to bear firearms in national parks. In this example, 6 revisions of the Credit CARD act are shown.

67% of the bill's sections are classified by our analysis into the category of Natural Resources, assisting the user in focusing in on either that portion of the bill or to the outlier topics.

EVALUATION

The examples described above demonstrate the way in which Many Bills can provide guidance towards understanding a bill's content. To understand how people actually use Many Bills's features and to evaluate whether or not we met our design goals, we analyzed web log data and conducted interviews with select users. For evaluation purposes, the site was instrumented to record single-click events in addition to page views, capturing each click on the minify, read-in-full, and color toggle buttons.

In the five months since launching the site (end of March to beginning of September, 2010), 11,412 visitors visited the site. Examining the logs, we found that 6,656 of these visitors clicked beyond the home page, and this group collectively performed 99,746 actions (clicks) on the site. While these are high numbers, looking at just the users who performed more than 100 actions on the site, the number of users drops to 125 (1.1% of total users). These active users collectively account for 49,264 actions (49.4% of total actions).

This power law distribution of activity is typical for any web community site, and it provides an indication that Many Bills is highly engaging for a certain type of citizen user and that most visitors use the site with less commitment. To explore this hypothesis, we examined the behavior of these 125 "power users," contrasted with the behavior of the 6531 "casual users," who browsed beyond the home page. The purpose of this segmentation is to determine if there are behavioral differences between these populations, beyond simply high versus low levels of activity.

Analysis of Power Users

In terms of visits to the site, power users average 2.3 visits each (median 1, stdev 2.58, max 15). Each visit, they perform an average of 267.6 actions on the site (median 208, stdev 226.1). 39.2% of power users return to the site more than once, contrasted with just 13.5% of casual users.

A surprise to us is that casual users and power users find Many Bills through similar means and search for similar terms on the site. Approximately a third of users find the site through a site associated with our company, a quarter find it through a search engine, 15% find it through a technology or visualization related blog, and 10% find it through a government-related discussion site or blog. We expected that power users would have found the site more through government-related channels, but evidence indicates there is no difference.

Similarly, search terms used on the site by these two groups fall within the same general categories. 18% of users search for a specific bill name, about 8% search for a congressperson, and 3-5% of searches are for geographic regions (states, cities, countries), company names, and government agencies. The remaining 60% are general keyword searches such as "health care," "foster care," and "transportation planning."

The real differences between the casual and the power users emerges in how they interact with bills. Figure 7 shows a comparison of casual versus power users.

For each site action, except for viewing a Congressperson's page, the power users are performing the actions at a significantly higher level, as measured by Chi-square tests ($p<0.001$ for each test). This higher level not surprising because, by definition, the power users are much more active. What is interesting about their activity though is that power users are drilling down into the details of the bill more than they zoom out, whereas the casual users are choosing to view the highest level summary views, over viewing its full text. This can be seen in that twice as many power users choose to read the bill versus minified (98% vs. 54%) and twice as many casual users minified a bill versus read a bill in full (12% vs. 6%),. Similarly, while saving a bill to one's "bill tray" on the site is a less common

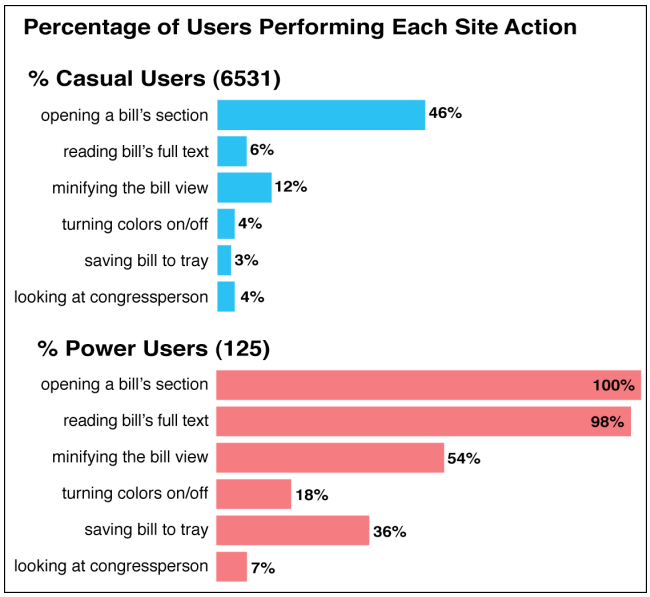

Figure. 7. Comparing all users to 125 power users in terms of how they spend their time on the site

activity, power users are doing it twice as much as they are turning the color guides on/off, whereas the casual users do both at the same low rate. This power-user action of saving a bill on the site indicates a longer-term intention to examine a bill and compare it with others.

Our conclusion from this is that power users, those most active on the site, are utilizing the features that enable the most detailed inspection of a bill: they are viewing bills in their entirety and saving them for later reference. In contrast, the casual users, who by definition are less committed to the site, use the minified feature more, which is designed for the quickest and most cursory overview of a bill's topics. We see this as evidence that our site and visualization design is able to support different levels of interest, and that the more engaged user is focused on accessing the details, rather than remaining at the overview level.

Interviews

To understand how Many Bills does, or does not, support these power users, we interviewed a small set of users who, in addition to having used the site, had experience with legislation in either a professional or personal capacity. We spoke with four people: a journalist, an employee of a non-profit focused on government transparency, a journalism professor, and a citizen involved his local political party chapter. We gave each user a short introduction to Many Bills and then directed them to the live system. We asked them to use the site as they wished for one week, to explore legislation that was of interest to them. We did not assign any specific tasks. One week later, we conducted a phone interview with each subject, asking them about their impressions and discoveries. The interviewees reported having spent from 30 minutes to over 2 hours on the site, spread over multiple days. Their comments, although not in direct comparison to other tools like THOMAS, were well-informed about legislation and legislative data available online.

Our interviews were conducted over the phone in a semi-structured way, gathering their impressions and reactions to the tool.

Overview of an Entire Bill

The primary message we heard during our interviews was that the subjects appreciated the approach on Many Bills of making the original text available, while providing a layer of abstraction that includes visual guides. For example, the non-profit worker said:

> *"Most people try to look at legislation in terms of summarizing it or news articles about it and what you guys got to the heart of was keeping the legislation intact, Keeping its original form but providing an overview insight without delving into summarizing it. It was a neat of experience of still having that first hand account of legislation in a format that was approachable"*
>
> -- Non-profit worker

The journalist described how, with the time constraints of his profession, having a tool that guides a journalist to the general topics is very useful:

> *"A lot of reporters don't have a lot of time anymore to go through and read a lot of legislation, sadly. And so this is a great way to give them an 'at a glance' at what some of the issues in the legislation are so that they know whether or not to dig a little bit deeper."*
>
> -- Journalist

The politically-active citizen was our most enthusiastic subject, explaining

> *"I don't know if it's my attention span or what it is, [but] when I look at the bills I glass out in the first paragraph. When I saw you took the bill and color-coded it and see what a paragraph was about, it suddenly made this very arcane area very approachable to me. I still had a lot to read but at least it was approachable."*
>
> -- Citizen

Many Bills does not eliminate the job of reading the bill, but, according to our subjects, the visualization made it easier to find on the portions of a bill that they were interested in reading and for getting an overall gist on the topics a bill covers.

Discovering Unusual or Outlier Sections

Because of our illustrative examples that motivated much of the design of the site, we were particularly interested in knowing how users found the text classification and color coding. All interview participants confirmed that the visual segmentation created by our coloring of sections helped them navigate the bills and make decisions about what to read next. Again, the journalist said:

"Finding stuff that's way outside the box was mostly what I was interested in. So I found myself looking more at those outliers. I'm glad that visualization [modes are] there because when you shrink it down and look at it to see what the biggest outliers are, that's a great tool to get a two-second glance at some thing that would take 3 days to read."

-- Journalist

The active citizen described a specific scenario in which the color-coding helped him. His local Democratic party interest group was interested in understanding a new energy bill's contents, S. 1733: Clean Energy Jobs and American Power Act. His plan had been to assign each person a fixed number of sections, to divide up the work of reading the bill and summarize it back to the group. "I tried to break it up to the group mechanically... The group didn't read the bill." When he first used Many Bills, he realized he could draw conclusions about the bill without the need for crowd-sourcing out the different sections:

"Everyone had a difficult time approaching the energy bill until we had it color-coded. I can begin to look at the sections in it.... What I did learn is that there are a lot of parts to a bill that have seemingly little to do with the substance that the bill is about. A lot is about how it will be instantiated and financially supported."

-- Citizen

This feedback that the color-coding gave users a way to locate sections that deviated from the bill's main topic has highlighted the importance of supporting the task of locating outlier sections within bills. In future work, we plan to offer more advanced analysis of outlier sections and enhance the ability for users to find them.

A New Perspective on Congresspeople

The congresspeople pages were added to the site a month prior to this writing, so there has been less time available for users to explore them. Our intention in designing this people-centric visualization was to provide an additional entry point to a bill that is personally meaningful to a user. By looking up one's own congressperson, a user can discover legislation that may be more relevant to their local region; by seeing what a congressperson has been actively sponsoring, a user can form new conclusions about the efforts of elected officials. For example, our active citizen used the feature to compare across multiple congresspeople:

"Looked at my congressperson. There were 177 bills, not many passed. Noticed he was very active... Looked at [House Majority Leader] Harry Reid . Noticed there were less bills... Looked at [House Minority Leader] Mitch McConnell. Saw two bills sponsored and five co-sponsored. It's an eye opener. This was opening up to me the inner workings of Congress. ... [Learned that McConnell's] job was not to represent the people of Kentucky; He was there to represent the Republican party.

-- Citizen

We also had an opportunity to give a demo of Many Bills to a Member of Congress. We showed her a bill that she had recently proposed relating to the BP oil spill in the Gulf of Mexico. While the record shows that the bill has not progressed through Congress, she explained that the contents of the bill were incorporated into another, more comprehensive bill that was voted on and passed. She elaborated that the process of legislation is very "human" and hard to fully capture in government documents. Our discussion with her highlighted a limitation of government transparency based on government-released data: the data does not tell the whole story. A major direction to take this research is to incorporate the input of expert users, through mechanisms like annotation, that explain the data and reveal the story behind the data.

IMPLICATIONS

As discussed earlier in the paper, our belief is that for the open government movement to engage citizens with the government data, tools for citizens need to provide guides and references to make finding the information easier, rather than simply summarize or provide interpretations of the data. In our power users, we see indications and reports that the detailed views into the data are the most utilized, and useful, capabilities. The implication is that open government data applications that do support data exploration, and essentially ownership of the data, empowers citizens and supports their goals.

On the other hand, we have not seen broad adoption on the site by casual users interested in government. This may be a consequence of our chosen dataset: legislative text is still a challenge for lay users to understand, even with topic guides. Yet this may be an issue with any open government dataset: the details, while important in terms of what the government is doing, may be beyond the scope of what citizens may want or need. This remains a challenge for those working in open government to continue to explore.

Our analysis of how people are using Many Bills provides some lessons learned that may apply to any applications built for large text datasets, either government-related or not. First, the activity on the site will likely follow a power law where a few number of people are engaging in great detail with the text, and these users have either a professional or deep personal interest in the data. That is to be expected for any resource of complex data. But what was informative in our log analysis and interviews was that the activity level of the user influenced which features they used: power users focused more on features for detailed analysis and casual users focused on features for summarizing analysis. This discovery indicates that the design of our data visualization, which aimed for a mid-level of detail as a default, was able to support *both* types of users by allowing one to either abstract or drill deeper on the data. This is not always the appropriate choice to make, but in our context of aiming to engage a diverse citizen population, we see this as a promising approach.

As we move forward in our development of open government applications aiming to engage citizens, we see allowing users, particularly power users, to annotate and

even contribute to the data, as an important step towards empowering citizens to engage with their government. We also see that involving power users in the process of explaining the data could open up channels through which the casual user may transition to being a more engaged user.

CONCLUSION

In this paper, we presented the design and implementation of Many Bills, a web-based set of visualizations that provides rich interactive views of US Congressional Legislation. Our initial user evaluations indicate that this is a valuable tool for helping people access the content of bills and get an easy to interpret overview.

The techniques used in Many Bills can be applied to domains with complex categorized documents that have an impact on non-experts. For example: legal documents such as terms of service and patent filings, other government documents such as judicial court decisions and the tax code. The approach used in Many Bills can assist users in accessing the content of the text with topic and outlier guides, pointing to the potentially interesting portions.

Many Bills is one contribution to the open government movement and there are many remaining opportunities in this domain. We see the road to citizen engagement as having many steps, beginning with citizens understanding the data and the actions of the government. This is where Many Bills' focus has been. Deeper engagement includes citizens communicating their discoveries and sharing their opinion. Many Bills supports some of this, but this could be supported to a greater extent.

ACKNOWLEDGMENTS

We want to particularly thank GovTrack.us and Open Congress for providing their data in open, accessible ways. We also thank our colleagues for their input on the design of Many Bills and our users for their valuable feedback.

REFERENCES

1. GovTrack.us. http://www.govtrack.us/
2. OpenCongress. http://www.opencongress.org/
3. THOMAS (Library of Congress). http://thomas.loc.gov/
4. Apitz, G. and Lin, J. Interfaces to Support the Scholarly Exploration of Text Collections. (2008).
5. Berkner, K., Schwartz, E.L., and Marle, C. Document Recognition and Retrieval XI. SPIE (2003), 54-65.
6. Boguraev, B., Kennedy, C., Bellamy, R., Brawer, S., Wong, Y., and Swartz, J. Dynamic presentation of document content for rapid on-line skimming. (1998), 118-128.
7. Collins, C., Viégas, F.B., and Wattenberg, M. Parallel tag clouds to explore and analyze faceted text corpora. IEEE Symposium on Visual Analytics Science and Technology (VAST), (2009), 91–98.
8. Don, A., Zheleva, E., Gregory, M., et al. Discovering interesting usage patterns in text collections: integrating text mining with visualization. (2007), 213–222.
9. Eick, S., Steffen, J., and Jr, E.S. Seesoft - A Tool for Visualizing Line Oriented Software Statistics. IEEE Transactions on Software Engineering 18, 11 (1992), 957-968.
10. Fekete, J. and Dufournaud, N. Compus: visualization and analysis of structured documents for understanding social life in the 16th century. ACM (2000), 47-55.
11. Havre, S., Hetzler, E., Whitney, P., and Nowell, L. ThemeRiver: Visualizing Thematic Changes in Large Document Collections. IEEE Transactions on Visualization and Computer Graphics 8, 1 (2002), 9-20.
12. Hearst, M.A. TileBars: visualization of term distribution information in full text information access. Proceedings of the SIGCHI conference on Human factors in computing systems, (1995), 59–66.
13. Keim, D.A. and Oelke, D. Literature fingerprinting: A new method for visual literary analysis. IEEE Symposium on Visual Analytics Science and Technology, 2007. VAST 2007, (2007), 115–122.
14. Kinnaird, P., Romero, M., and Abowd, G. Connect 2 congress: visual analytics for civic oversight. (2010), 2853–2862.
15. McCallum, A.K. MALLET: A Machine Learning for Language Toolkit. 2002. http://mallet.cs.umass.edu.
16. Nyhan, B. Why the "Death Panel" Myth Wouldn't Die: Misinformation in the Health Care Reform Debate. POLITICS 8, 1 (2010), 5.
17. Plaisant, C., Rose, J., Yu, B., et al. Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. ACM (2006), 141-150.
18. Stasko, J., G\örg, C., and Liu, Z. Jigsaw: Supporting investigative analysis through interactive visualization. Information Visualization 7, 2 (2008), 118–132.
19. Strobelt, H., Oelke, D., Rohrdantz, C., Stoffel, A., Keim, D.A., and Deussen, O. Document Cards: A Top Trumps Visualization for Documents. IEEE Transactions on Visualization and Computer Graphics 15, 6 (2009), 1145-1152.
20. Viégas, F.B., Wattenberg, M., and Feinberg, J. Participatory Visualization with Wordle. IEEE Transactions on Visualization and Computer Graphics 15, 6 (2009), 1137–1144.
21. Wattenberg, M. and Viégas, F.B. The word tree, an interactive visual concordance. IEEE Transactions on Visualization and Computer Graphics, (2008), 1221–1228.
22. Wise, J.A., Thomas, J.J., Pennock, K., et al. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. Proceedings of the 1995 IEEE Symposium on Information Visualization, (1995).