The Columbia

SCIENCE AND TECHNOLOGY LAW REVIEW

www.stlr.org

HACK, MASH, & PEER: CROWDSOURCING GOVERNMENT TRANSPARENCY

Jerry Brito[*]

In order to hold government accountable for its actions, citizens must know what those actions are. To that end, they must insist that government act openly and transparently to the greatest extent possible. In the twenty-first century, this entails making its data available online and easy to access. If government data is made available online in useful and flexible formats, citizens will be able to utilize modern Internet tools to shed light on government activities. Such tools include mashups, which highlight hidden connections between different data sets, and crowdsourcing, which makes light work of sifting through mountains of data by focusing thousands of eyes on a particular set of data.

Today, however, the state of government's online offerings is very sad indeed. Some nominally publicly available information is not online at all, and the data that is online is often not in useful formats. Government should be encouraged to release public information online in a structured, open, and searchable manner. To the extent that government does not modernize, however, we should hope that private third parties build unofficial databases and make these available in a useful form to the public.

[*] Senior Research Fellow, Mercatus Center at George Mason University. The author would like to thank Chris White and Mark Adams for their research assistance, as well as Jerry Ellig, Jim Harper, Tim Lee, and Kyle McKenzie for their valuable comments on early drafts of this Article. You may contact the author at gbrito@gmu.edu.

INTRODUCTION

The federal government makes an overwhelming amount of data publicly available each year. Laws ranging from the Administrative Procedure Act[1] to the Paperwork Reduction Act[2] require these disclosures in the name of transparency and accountability. However, the data is often only nominally publicly available. First, much government data is not available online or even in electronic format. Second, the data that can be found online is often not available in an easily accessible or searchable format. If government information were made public online in standard open formats, the online masses could help ensure the transparency and accountability that is the reason for making information public in the first place.

Part I of this Article will show that government information that is nominally publicly available is in fact difficult to access either because it is not online or, if it is online, because it is not available in useful and flexible formats. Part II explores how independent third parties have improvised where government has failed and made public information available online in flexible formats. Finally, Part III offers some recommendations for the government to improve its online offerings. It also argues that until such improvement takes place, private parties can fill the breach.

I. PUBLICLY AVAILABLE GOVERNMENT INFORMATION

Democracy is founded on the principle that the moral authority of government is derived from the consent of the governed.[3] That consent is not very meaningful, however, unless it is informed.[4] When a government makes decisions in secret, opportunity for corruption increases and accountability to the people decreases. This is why government transparency should be a priority. When official meetings are open to citizens and the press, when government finances are open to public scrutiny, and when laws and the procedures for making them are open to discussion, the actions of government enjoy greater legitimacy.

Recent years have seen a renewed effort to increase government transparency in the United States. In the wake of the Jack Abramoff,[5] Duke Cunningham,[6] and William

[1] Administrative Procedure Act §§ 3-4, 5 U.S.C. §§ 552-53 (2006).

[2] Paperwork Reduction Act, 44 U.S.C. §§ 3501-3520 (2006).

[3] *See* The Declaration of Independence para. 2 (U.S. 1776).

[4] The Framers, for example, required that Congress keep and publish a record of its activities. U.S. Const. art I, § 5, cl. 3 (“Each House shall keep a Journal of its Proceedings, and from time to time publish the same, excepting such Parts as may in their Judgment require Secrecy; and the Yeas and Nays of the Members of either House on any question shall, at the Desire of one fifth of those Present, be entered on the Journal.”).

[5] *See generally* Susan Schmidt & James V. Grimaldi, *Abramoff Pleads Guilty to 3 Counts*, Wash. Post, Jan. 4, 2006, at A1 (detailing the Justice Department’s “wide-ranging public

Jefferson[7] scandals, Congress has moved again to shed light on its own activities. In 2006, Senators Barack Obama and Tom Coburn introduced legislation requiring the full disclosure of all organizations receiving federal funds through an online database to be operated by the Office of Management and Budget (OMB).[8] The result was the Federal Funding Accountability and Transparency Act of 2006, which goes into full effect beginning January 1, 2008.[9] Additionally, House Democrats, led by Speaker Nancy Pelosi, pledged that after the 2006 congressional elections they would enact legislation to "restore accountability, honesty, and openness at all levels of government."[10] The result was the Honest Leadership and Open Government Act of 2007, which requires that information about earmarks be published on a public, searchable website 48 hours before a vote can be taken on the bill containing the earmarks.[11]

Laws encouraging government transparency and accountability have been a feature of the American system of government since the founding of the Republic. The Constitution, for example, requires that each house of Congress "keep a Journal of its Proceedings, and from time to time publish the same, excepting such Parts as may in their Judgment require Secrecy."[12] Today, the *Congressional Record* satisfies this requirement. With the advent of the regulatory state, the *Federal Register* was established to assemble a record of the actions of the new executive agencies.[13] The Administrative Procedure Act expanded the role of the publication by requiring agencies to publish not just their

corruption investigation" centered on former lobbyist Jack Abramoff).

[6] *See* Charles R. Babcock & Jonathan Weisman, *Congressman Admits Taking Bribes, Resigns*, Wash. Post, Nov. 29, 2005, at A1.

[7] *See* Jerry Markon & Allan Lengel, *Lawmaker Indicted on Corruption Charges*, Wash. Post, June 5, 2007, at A1.

[8] Press Release, Senator Barack Obama, Obama, Coburn Introduce Bill Requiring Public Disclosure of All Recipients of Federal Funding (Apr. 7, 2006), *available at* http://obama.senate.gov/press/060407-coburn_introduc/.

[9] Federal Funding Accountability and Transparency Act of 2006, Pub. L. No. 109-282, 120 Stat. 1186.

[10] Office of the House Democratic Leader Nancy Pelosi, 109th Congress, A New Direction for America 21 (2006), *available at* http://www.speaker.gov/pdf/thebook.pdf.

[11] Honest Leadership and Open Government Act of 2007, Pub. L. No. 110-28, § 521, 121 Stat. 735, 760-64.

[12] U.S. Const. art I, § 5, cl. 3.

[13] Office of the Fed. Register, Nat'l Archives and Records Admin., A Brief History Commemorating the 70th Anniversary of the Publication of the First Issue of the Federal Register 2-4 (2006), *available at* http://www.archives.gov/federal-register/the-federal-register/history.pdf.

orders, but their proposed rules and other documents as well.[14] More recently, the Freedom of Information Act for the first time gave Americans the right to access the general records of federal agencies.[15]

A. Public Government Data is Often Not Online

Unfortunately, many of the statutory requirements for disclosure do not take Internet technology into account. For example, the 1978 Ethics in Government Act requires the disclosure of financial information–including the source, type, and amount of income–by many federal employees, elected officials, and candidates for office, including the President and Vice President, and members of Congress.[16] The Act further requires that all filings be available to the public, subject to certain limited exceptions.[17] One might imagine, then, that every representative or senator's information would be just a web search away, but one would be wrong.

Members of the House of Representatives must file their disclosures with the Clerk of the House of Representatives, while Senators must do the same with the Secretary of the Senate.[18] Each of these offices maintains a searchable electronic database of the filings.[19] However, to access these databases, citizens must go to Washington, DC, and visit those Capitol Hill offices during business hours.[20] There are no other means of searching the databases, something that presents a major barrier to widespread dissemination of nominally publicly available information. Making such a database available to the public online can likely be accomplished at a negligible marginal cost given that the Clerk of the House and the Secretary of the Senate already have websites on which the information could be posted.

Outside of Congress, the President and Vice President, candidates for those offices, and other executive officials (including all Senate-confirmed officials) must file

[14] *Id.* at 6; 5 U.S.C. § 553(b) (2006).

[15] Freedom of Information Act, 5 U.S.C. § 552 (2006).

[16] Ethics in Government Act of 1978 §§ 101-02, 5 U.S.C. app. 4 §§ 101-02 (2006).

[17] 5 U.S.C. app. 4 § 105 (2006).

[18] 5 U.S.C. app. 4 § 103(h)(1) (2006).

[19] The Open House Project, Sunlight Found., Congressional Information & the Internet: A Collaborative Examination of the House of Representatives and Internet Technology 45 (2007) [hereinafter Open House Report], *available at* http://www.theopenhouseproject.com/report/openhouseproject_may8_07.pdf.

[20] *Id.*; *see also* Rob Bluey, *Why Aren't These Documents Available Online?*, The Open House Project, Mar. 7, 2007, http://www.theopenhouseproject.com/2007/03/07/why-arent-these-documents-available-online/ (describing the process of accessing documents at the House Legislative Resource Center).

their financial disclosure with the Office of Government Ethics,[21] which is specifically charged with making the filings available to the public.[22] However, there is no searchable database of these records available to the public. Instead, one must fill out and submit a form listing the persons whose disclosure forms one would like to view, and these are then copied and mailed.[23] Public access to an electronic relational database with this sort of information would allow for far more interesting uses, such as querying to see which sources of income appear most frequently (or contribute the most income overall) in the disclosures.

B. Online Public Government Data is Difficult to Use

Even when public information is available online, it is often not available in an easily accessible form. If data is difficult to search for and find, the effect might be the same as if it were not online. Additionally, to allow users to exploit the full potential of the Internet–to subscribe to data streams and to mix and match data sources–data must be presented in a structured machine-readable format.

For example, the Federal Communications Commission (FCC) is an independent government agency with an active regulatory agenda that it manages via its online docket system.[24] In theory, users of the FCC website are able to see active rulemakings, search for and read FCC documents and public interest comments filed by interested parties, and file their own comments. In practice, the site seems to be an exercise in obscurantism.

The main area of the FCC's home page contains a listing of news releases, commissioner statements, and public notices relating to new or existing regulatory proceedings.[25] These items are linked to both PDF and Microsoft Word files of the documents despite the fact that someone reading the page will be using a web browser, an application that generally reads neither of those formats. Accessing these documents requires launching a new application; and linking to a document–for example, linking to a commissioner statement from a blog entry–is less straightforward than linking to a simple web page.[26] In most cases, the documents listed on the home page pertain to an

[21] 5 U.S.C. app. 4 § 103(c) (2006).

[22] 5 U.S.C. app. 4 § 103(d) (2006).

[23] U.S. Office of Gov't Ethics, Request to Inspect or Receive Copies of SF 278 Executive Branch Personnel Public Financial Disclosure Reports or Other Covered Records - OGE Form 201 (2006), *available at* http://www.usoge.gov/pages/forms_pubs_otherdocs/fpo_files/forms/fr201_06.pdf.

[24] Federal Communications Commission, FCC Electronic Comment Filing System, http://www.fcc.gov/cgb/ecfs/ (last visited Oct. 10, 2007).

[25] *See* Federal Communications Commission, FCC Home Page, http://www.fcc.gov/ (last visited Oct. 10, 2007).

[26] The FCC Web site is fundamentally at odds with the ease of accessibility for which the

open regulatory proceeding, but there are no links to the docket where one could read public interest comments or other related documents.

The dockets containing proposed rules and other official FCC documents, as well as public comments, are available on the website through a search form.[27] There is neither an index of open proceedings nor indexes of documents within each proceeding docket. To obtain a listing of documents in a given docket, you must know the docket's number and search using that number. The resulting list is presented in chronological order with no way to sort by author, document length, or any other field. Additionally, there is no way of searching within dockets for specific keywords.[28] Even if there were a function that allowed one to search within documents, the results would be incomplete since many documents are posted as image files that are not easily parsed by computers and would not be returned in a search.[29] This applies both to comments submitted by the

World Wide Web was created. In the original document proposing a World Wide Web, Tim Berners-Lee and Robert Calliau specifically rejected the notion of "forc[ing] users to use any particular word processor, or mark-up format." Tim Berners-Lee & Robert Calliau, WorldWideWeb: Proposal for a HyperText Project, http://www.w3.org/Proposal.html (last visited Oct. 10, 2007). The genius of the Web is that anything that can be expressed in text can be published as an HTML document that can be read by anyone with a browser and that, more importantly, can be easily linked to and referenced by other HTML documents. In contrast, the FCC site breaks the conceptual model of an interlinked web and forces those who would link to information on the FCC site to place warnings that the linked-to information is not a webpage. *See, e.g.*, Carlo Longino, *FCC Says Rural Telcos Have to Play Nice With VoIP*, Techdirt, Mar. 2, 2007, 15:18 PST, http://www.techdirt.com/articles/20070302/073711.shtml (alerting readers that a linked document on the FCC website is a PDF).

[27] Federal Communications Commission, Electronic Comment Filing System, http://fjallfoss.fcc.gov/prod/ecfs/comsrch_v2.cgi (last visited Oct. 10, 2007). A note at the bottom of this page states that it was last updated on Dec. 11, 2003. *Id.*

[28] *See id.*

[29] One of the main document types used by the FCC is Portable Document Format (PDF). PDFs can contain digital text that is subject to search (usually created by saving as a PDF document from a word processing application) or images of text that cannot be searched (usually created by simply scanning a printed document). *See* Adobe Systems Inc., Adobe Reader 7.0 for Windows and Macintosh 166 (2004), *available at* http://www.adobe.com/products/acrobat/pdfs/acrruserguide.pdf ("PDF documents that are created by scanning a printed page are inherently inaccessible because the document is an image, not text that can be tagged into a logical document structure or reading order."). The Department of Justice (DOJ) has taken note of this problem as it relates to the accessibility of websites for disabled users who rely on devices that depend on machine-readable text. In a 2004 report, the DOJ stated:

> A more significant problem involves agencies' use of inaccessible content on their sites. An agency may create a Web page that is easily navigated by people using a text-only browser but then include downloadable files that are inherently inaccessible. This problem occurs most frequently with two types of file content used by many

public[30] and FCC documents.[31] This is the case even though public comments are usually created in word processing applications, such as Microsoft Word, which produce machine-readable electronic documents.

How do these non-searchable image files of documents come to be? First, if a commenter opts not to submit their comment electronically and instead mails or sends by courier a physical copy of the document, the FCC scans the document as an image.[32] Second, even if a commenter is submitting a comment electronically, the commenter has the option to submit a non-searchable image file. The result, as one commentator put it, is that "[t]here is no incentive for filing parties to make their documents machine readable and they may prefer to make them difficult to use in order to increase the burden on opposing parties filing reply comments under short deadlines."[33]

Some agencies, such as the EPA, do not house their electronic dockets on their own websites. Instead they use Regulations.gov, a combined federal regulatory docket system managed by the Office of Management and Budget and part of President George W. Bush's "eRulemaking Initiative."[34] Acknowledging that "online access to comments

> components—files rendered by scanning to Adobe Acrobat's portable document format (pdf) and multimedia files.

Department of Justice, Information Technology and People with Disabilities: The Current State of Federal Accessibility § III.1 (2000), *available at* http://www.usdoj.gov/crt/508/report/content.htm (last visited Oct. 10, 2007). Revised DOJ guidance suggests that agencies may use PDF files as long as they take care to ensure that they are accessible. Department of Justice, Section 508 of the Rehabilitation Act: Accessibility for People with Disabilities in the Information Age (Results of 2001 Survey) § II, *available at* http://www.usdoj.gov/crt/508/report2/web.htm (last visited April 17, 2008).

[30] *See, e.g.*, Comments of Verizon Wireless to the FCC, In re: Auction of 700 MHz Band Licenses Scheduled for January 16, 2008 (Aug. 31, 2007), *available at* http://fjallfoss.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6519721231 (illustrating use of PDF documents for comments submitted to FCC).

[31] *See, e.g.*, Federal Communications Commission, Auction of 700 MHz Band Licenses Scheduled for January 16, 2008, Comment Sought on Competitive Bidding Procedures for Auction 73 (Aug 17, 2007), *available at* http://fjallfoss.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6519611785 (illustrating use of PDF documents for official FCC documents).

[32] This does not explain however why the FCC's own documents, such as the public notice cited in footnote 31, are also scanned as images. Nor does it explain why optical character recognition is not generally applied to scanned documents.

[33] Michael Marcus, *FCC Website: The good, the bad, and the ugly*, Spectrum Talk, May 16, 2006, 08:26 EDT, http://spectrumtalk.blogspot.com/2006_05_01_archive.html.

[34] Office of Management and Budget, Report to Congress on the Benefits of the President's E-Government Initiatives 4-5 (2007), *available at* http://www.whitehouse.gov/omb/egov/documents/FY07_Benefits_Report.pdf. While OMB is

about regulations, along with other supporting documents, is limited," the Bush initiative sought to ease matters by creating one website at which users could find, read, and comment on regulations.[35] The Regulations.gov site currently provides the ability to search and view all rulemaking documents published in the *Federal Register*, and to submit comments to some agencies on their open proceedings.[36] It also houses the complete dockets (i.e., all notices, technical reports, and public comments) of over thirty participating agencies, including the Environmental Protection Agency.[37] The initiative's objective is to eventually house all federal dockets in one unified "Federal Docket Management System."[38]

Unfortunately, the site leaves much to be desired.[39] Like the FCC's site, Regulations.gov does not offer a full text search of documents.[40] Users can only search by titles, authors, and other limited fields.[41] As one law librarian, Barbara Brandon of the University of Miami School of Law, explained this shortcoming:

> "If I wanted to find all comments made by the Natural Resources Defense Council in an EPA rulemaking docket, for example, I would put in their full name or 'NRDC' and the Web site might give me something like 30 hits showing their comments."
>
> However, what the search would not find, Brandon explained–and what is of particular interest to outside advocacy groups and researchers–are all the documents within a docket that happen to mention NRDC.

responsible for all of the President's e-government initiative, the EPA is the managing partner in charge of Regulations.gov. *Id.*

[35] *E-Government Reauthorization Act: Hearing on S. 2321 Before the S. Comm. on Homeland Security and Government Affairs*, 110th Cong. (2007) (statement of Karen Evans, Administrator of the Office of Electronic Government and Information Technology, OMB); *available at* http://hsgac.senate.gov/_files/121107Evans.pdf; *See also* Office of Management and Budget, Presidential Initiatives: E-Rulemaking, http://www.whitehouse.gov/omb/egov/c-3-1-er.html (last visited Oct. 10, 2007) (describing the EPA's eRulemaking Program).

[36] *See* Establishment of a New System of Records Notice for the Federal Docket Management System, 70 Fed. Reg. 15,086 (proposed Mar. 24, 2005) (effective May 3, 2005).

[37] *Id.*

[38] *Id.*

[39] *See* Ralph Lindeman, *Structural, Other Flaws Said to Impede Effectiveness of E-Rulemaking Web Site*, BNA Daily Report for Executives, Mar. 30, 2007, at C-5; Cindy Skrzycki, *Document Portal Sticks on Funding*, Wash. Post, Jan. 10, 2006, at D1.

[40] Lindeman, *supra* note 39.

[41] *Id.*

> "In other words, with a full text search, I would be able to see all the agency documents that cite NRDC's comments, which could tell me what the agency had to say about the comments," Brandon said. "I would also be able to see what other commenters said about NRDC's comments."[42]

Twenty-seven federal agencies have migrated their dockets to Regulations.gov, which according to OMB accounts for eighty-two percent of all federal regulations.[43] While this process toward centralization has been hailed as a success, it may in fact be a disaster. While efficient in theory, consolidation may be a step backward if the centralized database does more to obscure data than to make it easily accessible. A few days after Regulations.gov won an award from *Government Computer News*,[44] the Congressional Research Service (CRS) issued a report outlining serious questions regarding the site, including "the general navigability of the website, the consistency and completeness of the data, [and] whether the system allows users to adequately search existing dockets."[45] The report catalogues several attempts by CRS to find information using the site's navigation or search functions. These search attempts were unsuccessful and yielded thoroughly confusing results.[46]

C. The Promise of Structured Data

Neither the FCC website nor Regulations.gov offer access to their data in a structured format. What does this mean? The most common form of subscribable structured data is an RSS feed. RSS stands for "really simple syndication" and usually refers to a family of data formats that allow the automation and aggregation of data.[47] For example, the *New York Times* offers an RSS feed for its homepage,[48] as does the

[42] *Id.*

[43] Rutrell Yasin, *Agency Award—Environmental Protection Agency*, Government Computer News, Oct. 8, 2007, *available at* http://www.gcn.com/print/26_26/45188-1.html.

[44] *Id.*

[45] Curtis W. Copeland, Congressional Research Service, *CRS Report for Congress: Electronic Rulemaking in the Federal Government* 35-39 (2007), *available at* http://www.opencrs.com/document/RL34210.

[46] *Id.*

[47] *See* Mark Pilgrim, *What is RSS*, O'Reilly XML.com, Dec. 18, 2002, http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html.

[48] XML feed for items on the *New York Times* home page, http://www.nytimes.com/services/xml/rss/nyt/HomePage.xml (last visited Oct. 10, 2007).

Washington Post.[49] A user can subscribe to these feeds with a desktop application called a "feed reader"[50] or a web-based reader such as Google Reader.[51] Any time something is added to the home page of the newspaper, it is simultaneously published in that newspaper's RSS feed. When subscribers turn on their feed reader, it checks all the subscribed feeds for new items, which are then displayed. So, with one simple feed reader application, a user can keep track of dozens or hundreds of feeds without having to regularly visit the websites of the publisher, in this case the newspapers.

Imagine being able to subscribe to feeds from Regulations.gov or individual agency websites. Subscribe to the FCC's RSS feed and then never visit the site again just to check if new regulations have been proposed.[52] If a new regulation (or other document) is added your reader automatically alerts you.[53] But it could be even more useful. The *New York Times*, for example, offers a feed just for its automotive section.[54] Subscribe to it and you are notified only when new articles about cars are published and you never have to wade through all the other content the *Times* publishes. The *Times* also offers feeds for its food section, its "Europe news" section, and dozens more. There is no reason why the FCC could not similarly publish a feed for each of its bureaus to be subscribed to by persons interested just in wireless spectrum regulations or just cable regulations.

Giving individual users such capabilities would greatly increase transparency. More importantly, however, structured data formats such as RSS offer even greater potential for openness because they make data more accessible and flexible. Once a user is aware of a regulation they would like to track, why not allow them to "subscribe" to the regulation? Blogs are preeminent users of structured data and the vast majority offer RSS feeds to which you can subscribe. Subscribe to a dozen and whenever you turn on

[49] XML feed for items on the *Washington Post* front page, http://feeds.washingtonpost.com/wp-dyn/rss/print/index_xml (last visited Oct. 10, 2007).

[50] *See* Ellen Finkelstein, Syndicating Web Sites with RSS Feed for Dummies 10-11 (2005).

[51] Google Reader, http://reader.google.com (last visited Oct. 10, 2007).

[52] There are scattered uses of RSS in government. The SEC for example provides RSS feeds for financial disclosure data that firms voluntarily submit in a structured format. *See* SEC XBRL RSS Feed Files, http://www.sec.gov/info/edgar/ednews/xbrlrss.htm (last visited Oct. 10, 2007). The Copyright Office uses RSS extensively for its publications, including its Federal Register notices. *See* Copyright Office Federal Register Notices, http://www.copyright.gov/fedreg (last visited Oct. 10, 2007).

[53] Regulations.gov does offer e-mail alerts. Regulations.gov User's Guide § 9, http://www.regulations.gov/fdmspublic/help/en/PublicHelpGuide/PublicHelpGuide.htm#9_Notifications.htm (last visited Oct. 10, 2007). The site will e-mail you each time a new document is added to a particular docket to which you subscribe. However, complete dockets including public comments are available only for a few participating agencies.

[54] XML feed for items on the automotive section of the *New York Times* website, http://www.nytimes.com/services/xml/rss/nyt/Automobiles.xml (last visited Oct. 10, 2007).

your reader, the latest postings from each blog are available for you to read. Additionally, most blogs allow readers to leave comments at the end of blog posts, and many allow readers to subscribe to those comments. For example, suppose you subscribe to an automotive blog that you read regularly using your feed reader. One day the blog features a post about the recall of a brand of tires that you own. You know a bit about the issue, so you post a comment to the blog and you then want to track responses to your comment and the general conversation that develops on the blog around the tire recall issue. So you subscribe to the RSS feed for that particular post's comments and each time you turn on your reader you get not just the latest posts to the blogs you follow, but the comments posted by readers to the blog posts you're tracking. There is no reason why a government website could not allow users to subscribe to regulatory dockets and be notified of all official actions and public interest comments filed in a particular docket.

The *New York Times* also offers a series of "Times Topics" web pages and companion RSS feeds.[55] These range from persons (Rupert Murdoch, Hillary Clinton) to countries (Sudan, Colombia) to organizations, general subjects, and issues (New York Yankees, Supreme Court, United States, cancer). Subscribe to the RSS feed for one of these keywords and your feed reader will display articles published relating to that keyword anywhere in the pages of the *Times*. Imagine if such keyword subscriptions were available from regulatory agencies. The EPA, for example, could offer topic subscriptions such as "pesticides," "superfund," or "Vermont," making it easier for citizens to engage in the topics that matter to them.

Finally, even if the government cannot predict every possibly useful topic, readily available technology today allows for RSS subscriptions to keyword searches. Google News, for example, allows users to make a regular web search and then to subscribe to the results.[56] Each time a new item using your search term appears anywhere on the web, you are alerted.

II. Making Government Data Available and Useful Online

Making government information available online would not only benefit individual users of government websites, it would also make it simpler for third parties to aggregate government data. By aggregating data, websites can present government information in innovative and useful ways. For example, federal spending data gathered from a government website could be presented by a third party as an interactive map that shows the locations of funding recipients.[57] Such applications make data exponentially

[55] *New York Times* Times Topics, http://topics.nytimes.com/top/reference/timestopics/ (last visited Oct. 10, 2007).

[56] Google News, Browse Google News Help—About Feeds, http://www.google.com/support/news/bin/answer.py?hl=en&answer=59255 (last visited Oct. 10, 2007).

[57] Sunlight Foundation, Earmark Map, http://sunlightlabs.com/earmarks/ (last visited Oct. 10, 2007).

more valuable. Government need not develop such innovative tools itself; as long as the data is made available online in a structured format, private parties will make good use of it.

As we have seen, "structured data" is a term of art. It means that information is presented in a format that allows computers to easily parse and manipulate it. While a static web page that lists a series of news stories or proposed regulations is not structured, the web page may have a companion XML file containing the same information. A structured XML file would allow a user to sort the data by ascending or descending date, alphabetically by headline or author, by number of words, and in many other ways that a static web page does not afford.

In 2007, a group of interested citizens collaboratively produced a report detailing how the House of Representatives could use Internet technology to better serve its constituents. In it they explained,

> The notion of structured data is not new to the federal government. The Census Bureau, for instance, has for years not only provided a Web interface for census statistics–that is, a page where users can find simple data such as population numbers–but also the complete set of numeric data files to be downloaded and imported into database and statistics programs. The benefit of a download of the data is that with the complete data set computers can help people delve more deeply into the data and put it in new forms, such as charts and maps, that would be too time consuming to create by hand. Another example is the Securities and Exchange Commission's practice of making investment filings available to the public in XML format through its EDGAR program. Likewise, the Federal Election Commission makes campaign contribution information available in a downloadable structured data format, allowing the public to absorb the information in a variety of ways.[58]

When the government makes data available in a structured format, it opens the doors to innovative and enlightening remixes of information known as mashups. Mashups are tools that can potentially be used by journalists, bloggers, and citizens–the Internet's intelligent crowds–to better scrutinize government's activities. When government does not make data available online, or makes it available but not in a structured format, third parties take it upon themselves to fill the void by implementing ingenious hacks.

A. Hacks

Because of how the popular press has used it, the word "hack" is often misunderstood to mean only illicit access to computer networks. In fact, in tech circles that is only one possible meaning. Usually the term means "a clever or quick fix to a computer program problem," and also, "a modification of a program or device to give the

[58] Open House Report, *supra* note 19 at 11 (citations omitted).

user access to features that were otherwise unavailable to them."[59] It is this latter definition that is relevant here.

A number of independent third parties have created hacks that make available online, in a structured format, data that the government has either not put online or not made easily accessible. For example, disclosure forms for members of Congress are available online from *The Washington Post*'s U.S. Congress Votes Database.[60] Using this database, a user can look up a page for any member of Congress. The page includes a photo, a short biographical sketch, voting record, and much more information, including links to the past two years' financial disclosure forms. Where does the *Post* get this data? For House members, the Office of the Clerk once a year makes available electronically all the disclosure forms on a CD-ROM.[61] The *Post* uses this data to populate its online database.[62] For Senators, however, the *Post* must acquire physical copies of the filings and then scan them in order to make electronic copies.[63] While government has failed to provide the data online and requires citizens to make a formal request for physical copies of these public documents, the *Washington Post's* hack offers easy online access.[64]

Another independent third party that is hacking government data to make it accessible and useful to the public is GovTrack.us, a website by linguistics graduate student Joshua Tauberer.[65] GovTrack.us attempts to overcome the poor formatting of legislative information made available by the government. By scouring disparate and obscure government sources of congressional data, the site is able to create a unified and structured information resource. As Tauberer explains:

> The site is possible because the government has been posting the relevant information online for a while, but in scattered locations. For instance, legislation is posted in one place and votes on the very same legislation in another. . . .
>
> Each day GovTrack screen-scrapes these sites to gather the new

[59] Wikipedia, Hack (technology), http://en.wikipedia.org/wiki/Hack_(technology) (last visited Oct. 10, 2007).

[60] The Washington Post, The U.S. Congress Votes Database, http://projects.washingtonpost.com/congress/ (last visited Oct. 10, 2007).

[61] E-mail from Derek Willis, Research Database Editor, *The Washington Post*, to Mark Adams, research assistant to the author (Sept. 8, 2007, 09:47 EST) (on file with author).

[62] *Id.*

[63] *Id.*

[64] Unfortunately, the database only extends to members of Congress and not all federal employees who are subject to financial disclosure requirements.

[65] GovTrack.us: Tracking the U.S. Congress, http://www.govtrack.us/ (last visited Oct. 10, 2007).

> information. The information gets normalized and goes into XML files.[66]

What Tauberer is referring to are the Library of Congress's THOMAS online legislative database, and the House and Senate's practice of publishing daily roll call votes on their websites.[67] Individually, these data sources are certainly useful, but they do not come close to meeting the potential of modern technology.

For example, THOMAS is the go-to source for bills before Congress. The THOMAS website includes full text search of bills, their status, sponsors, committee reports, and other information.[68] However, pages on the THOMAS site use temporary web addresses that expire after a few minutes, making it difficult for users to bookmark, email, or otherwise share information.[69] Also, the legislative information offered by THOMAS is not available in a structured format. While one can find a list of all bills sponsored by a particular member of Congress on THOMAS, one cannot subscribe to a feed in order to be alerted whenever a particular member introduces new legislation.[70] Nor can one subscribe to a particular bill and be alerted to any actions related to it.

As we have seen, both the House and the Senate publish the results of daily roll call votes on their websites. The House publishes the data in a standard structured format known as XML. The Senate does not.[71] In neither case is it possible for a user to look up the voting record of a particular member of Congress. Both the House and the Senate present a web page for each bill considered and then list the names of all those voting yea, then all those voting nay, and finally those not voting.[72] To see a particular senator's complete voting record for the year, one would have to click on hundreds of pages and record whether the senator was listed as voting yes or no for each individual bill.

In contrast, Tauberer's GovTrack.us, as well as *The Washington Post*'s U.S. Congress Votes database, display complete voting records for individual members. Additionally, because GovTrack.us and U.S. Congress Votes present data in a structured

[66] Joshua Tauberer, *GovTrack.us, Public Data, and the Semantic Web*, O'Reilly XML.com, Feb. 8, 2006, http://www.xml.com/pub/a/2006/02/08/govtrack-us-public-data-semantic-web.html.

[67] *Id.*

[68] The Library of Congress, About THOMAS, http://www.thomas.gov/home/abt_thom.html (last visited Feb. 23, 2008).

[69] Open House Report, *supra* note 19, at 10.

[70] *See generally* The Library of Congress, THOMAS Home, http://thomas.loc.gov (last visited Feb. 23, 2008) (see "Browse Bill by Sponsor" dropdown menu).

[71] *See, e.g.*, Office of the Clerk, U.S. House of Representatives, Final Vote Results for House Roll Call 864, http://clerk.house.gov/evs/2007/roll864.xml (last visited Feb. 23, 2008) (roll call vote data available in XML).

[72] *See id.*; *see also* U.S. Senate, U.S. Senate Roll Call Votes 110th Congress - 1st Session, http://www.senate.gov/legislative/LIS/roll_call_lists/roll_call_vote_cfm.cfm?congress=110&session=1&vote=00372 (last visited Feb. 23, 2008) (roll call vote data unavailable in XML).

format, one can subscribe to feeds for members, and be notified daily via e-mail or RSS as to how these members voted on bills.[73] Other features these sites provide include

- subscribing to a bill and being alerted to every change or action (amendments, related hearings, votes, etc.)[74]
- subscribing to a member and being alerted not just to votes, but also to bills introduced and speeches made by that member[75]
- statistical facts for individual members, including the percentage of votes missed and the number of times the member has voted with and against the majority of his party[76]

To make these features possible, GovTrack.us and *The Washington Post* had to hack the data provided by Congress into useful structured formats. Because the House already provides roll call vote data in a structured XML format, it is easy for the sites to download the information from the House site and parse it into their own databases.[77] Senate vote data and THOMAS legislative information, however, are not available in a structured format,[78] so they must instead be "screen-scraped."[79]

In essence, "screen-scraping" involves calling up the web page that displays the

[73] The Washington Post, The U.S. Congress Votes Database: RSS Feeds, http://projects.washingtonpost.com/congress/rss (last visited Feb. 23, 2008); GovTrack: How To Use Trackers, http://www.govtrack.us/users/aboutmonitors.xpd (last visited Oct. 10, 2007).

[74] The U.S. Congress Votes Database, *supra* at 73; GovTrack, *supra* at 73.

[75] The U.S. Congress Votes Database, *supra* at 73; *see* GovTrack, *supra* at 73.

[76] *See, e.g.*, The Washington Post, The U.S. Congress Votes Database: Rep. Neil Abercrombie, http://projects.washingtonpost.com/congress/members/a000014 (last visited Feb. 23, 2008) (illustrating the information available for individual members).

[77] E-mail from Derek Willis, Research Database Editor, *The Washington Post*, to Mark Adams, research assistant to the author (Sept. 8, 2007, 09:47 EST) (on file with author) ("You can do a View Source on the House vote and see the XML. We use a Python XML parser to do the work there. For the Senate, we do indeed screen-scrape the vote using pattern-matching (also called regular expressions)."); *see also* Tauberer, *supra* note 66 (describing the process used by GovTrack to create and manipulate XML).

[78] Tauberer, *supra* note 66; Willis, *supra* note 77.

[79] Wikipedia, Screen Scraping, http://en.wikipedia.org/wiki/Screen_scraping (last visited Feb. 25, 2008) ("Screen scraping is a technique in which a computer program extracts data from the display output of another program. . . . The key element that distinguishes screen scraping from regular parsing is that the output being scraped was intended for final display to a human user, rather than as input to another program, and is therefore usually neither documented nor structured for convenient parsing.").

type of data the user wishes to gather (for example, a senate roll call vote page), identifying the patterns apparent on the page (such as where the bill title and number are displayed and which boxes correspond to the yeas and nays), and then writing a computer script that will transfer data found in designated display positions to the appropriate fields in a database.[80] In many ways this is the digital equivalent of having to scan paper copies of documents because, while the original may well be electronic in this case, it is the final user display that is accessed and parsed into meaningful groupings. In short, it is an inefficient and often inexact method.

Websites such as GovTrack.us and the U.S. Congress Votes Database must nevertheless be commended for cleverly employing "screen scraping" in order to bring the public better access to public information. Government could make this costly maneuvering unnecessary by providing data in a structured format.

So why is it that the government very often does not provide online data in useful forms? In most cases it is likely the result of bureaucratic inertia. In others, however, it must be noted that those in government have no incentive, and often a disincentive, to make public information easily accessible.

Derek Willis, one of the creators of *The Washington Post*'s congressional database, discovered during the course of his research that the Senate had experimented with publishing XML files of vote data for past sessions on its website.[81] This discovery demonstrated that the Senate has the ability to make its votes available online in a structured format. Willis wrote to the Senate Webmaster asking if structured voting data was available for the current session and, if so, would this data be made public.[82] The telling response read in part:

> A few representative votes (only a few from the early congresses) were published out to the active site during some testing periods. I really need to remove them from the site.
>
> We are not authorized to publish the XML structured vote information. The Committee on Rules and Administration has authorized us to publish vote tally information in HTML format [not a structured format]. Senators prefer to be the ones to publish their own voting records. As you know, looking at a series of vote results by Senator or by subject does not tell the whole story. Senators have a right to present and comment on their votes to their constituents in the manner they prefer. This issue was reviewed again recently and the policy did not change.[83]

[80] Tauberer, *supra* note 66; Willis, *supra* note 77.

[81] E-mail from Derek Willis, Research Database Editor, *The Washington Post*, to Cheri Allen, Senate Webmaster (Nov. 1, 2007, 18:07 EST) (on file with author).

[82] *Id.*

[83] E-mail from Cheri Allen, Senate Webmaster, to Derek Willis, Research Database Editor, *The Washington Post* (Nov. 16, 2007, 15:42 EST) (on file with author).

Senators doubtlessly would "prefer to be the ones to publish their own voting records."[84] But jealous control over information by government is anathema to democracy. Looking at a series of votes by a senator does in fact tell the "whole story" of that senator's voting record, and despite what the Webmaster may say, senators do not have a "right" to present their votes to the public "in the manner they prefer."

Other independent websites that are hacking government data to make it more useful and accessible to the public include:

- **LOUIS**–The Library of Unified Information Sources is a search engine that indexes Congressional Reports, the *Congressional Record*, congressional hearings, the Federal Register, presidential documents, GAO reports, and congressional bills and resolutions. The site allows users to subscribe to a search query via RSS so that they are alerted each time a new document references their search terms.[85]

- **MetaVid**–While Congress often streams live video feeds of committee hearings and other proceedings, these videos are rarely archived and simply disappear into the ether once the broadcasted event concludes.[86] MetaVid is a site that captures and archives these videos and makes them available to the public.[87] By capturing closed-caption data from cable TV broadcasts, MetaVid is able to make videos searchable by keyword.

- **OpenSecrets.org**–This comprehensive website by the Center for Responsive Politics gives users access to several informative databases. First, it takes campaign finance data available electronically from the Federal Election Commission and creates a database that allows users to search contributions, donors, candidates, and committees. Beyond this, however, the site groups contributions by industry sector (insurance, education, tobacco, etc.). Second, the site provides online access to the financial disclosure forms of Congress members and many executive officers. The site parses the data in these forms to produce a database that reveals the most popular financial investments of the disclosing parties, the top and bottom earners, and other relevant figures. Third, the site gathers lobbyist registrations, as required by the Lobbying Disclosure Act of 1995, from the Senate website and creates a database that allows users to search this data by

[84] *Id.*

[85] Sunlight Foundation, LOUIS – the Library of Unified Information Sources, http://www.louisdb.org/ (last visited Feb. 23, 2008).

[86] Open House Report, *supra* note 19, at 51-54.

[87] About MetaVid, http://metavid.ucsc.edu/wiki/index.php/About (last visited Apr. 6, 2008).

> client, lobbyist, industry, issue, agency, or bill.[88]

The most important contribution all these hacks make, however, may not be the accessibility they provide to individual users, but the fact that their hacked data is offered in a structured and open format. This allows yet other third parties to tap into the now useful data and create new applications. As Joshua Tauberer has explained, "Gathering the information in one place and in a common format gives rise to new ways of mixing the information together."[89]

B. Mashups

The value of structured data extends beyond the revolutionary accessibility it can provide individuals. Perhaps more importantly, it enables a class of applications known as "mashups" that combine two or more sets of data, resulting in a novel source of information.

The term "mashup" has its origins in music. The advent of digital editing technologies made it relatively simple for DJs and amateurs to take two or more different songs and mash them together to create novel creations.[90] The paradigmatic example of a music mashup may be Danger Mouse's highly acclaimed and highly illegal "The Grey Album," which mixed music from The Beatle's "The White Album" with vocals from rapper Jay-Z's "The Black Album."[91]

The term mashup now extends to applications that mix together disparate sets of data to create new and unique information.[92] For example, the popular free classified ad site CraigsList.com is an almost definitive source for rental housing listings in urban areas. However, the site lists ads in the order that users add them to the site. This means that, using the Washington D.C. metro area as an example, one listing could be for an apartment in the Adams Morgan neighborhood of the District and the very next ad would be for a house in Arlington, Virginia. This frustrated software engineer Paul Rademacher

[88] *See* OpenSecrets.org, Money in Politics Data, http://www.opensecrets.org/ (last visited Feb. 23, 2008).

[89] Tauberer, *supra* note 66.

[90] *See generally* Wikipedia, Mashup (music), http://en.wikipedia.org/wiki/Mashup_(music) (last visited March 5, 2008) ("A mashup . . . or bootleg is a song or composition created from the combination of the music of one song with the a cappella from another.").

[91] *See* Wikipedia, The Grey Album, http://en.wikipedia.org/wiki/The_Grey_Album (last visited March 5, 2008) ("The album quickly became extremely popular and well-distributed over the internet because of the surrounding publicity. It also came to the attention of the critical establishment; it received a very positive write-up in the February 9, 2004 issue of *The New Yorker* and was named the best album of 2004 by *Entertainment Weekly*.").

[92] *See* Wikipedia, Mashup (web application hybrid), http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid) (last visited Feb. 17, 2008).

when he was looking for a place to live in Silicon Valley in 2004.[93] So, he built HousingMaps.com, a mashup of the listings from CraigsList.com and Google Maps. This mashup allows users to bring up a map of the area in which they are interested (say five square blocks in a particular neighborhood). Then, pushpin icons will appear representing the properties available for rent in that area. Clicking on a HousingMaps.com pushpin brings up a bubble with the rental listing data including rooms, price, location, photos and a link to the actual listing.

What is amazing about a service like HousingMaps.com is that it is a new and unique information source that is richer and more useful than either craigslist or Google Maps alone. What makes this possible is Google's choice to make its maps application interface open for anyone to use and Craigslist's similar choice to make its data freely available in an open and structured format. These decisions to support openness and useful data formats allowed for an innovation that neither company could have predicted would emerge.

Indeed, when a site makes its data available in open formats, it cannot conceive of the many creative ways the data will be put to use.[94] Book Burro, for example, is a plug-in for the popular Firefox web browser that senses when you are looking at a page for a book (at Amazon.com, for example) and then fetches and displays data about the book's availability at local libraries, as well as prices at competing online stores.[95] Another example, Trendio, uses open application interfaces from Yahoo, Google, and Technorati to index articles emanating from over 3,000 news sources. It tracks the relative trends for words contained in those articles. The result is an index of trends in the media.[96]

Mashups built on open interfaces and structured data represent a great potential fount of information about the workings of government. Mashups produce varied and unexpected outcomes that could make government activities more transparent, and reveal patterns currently hidden in murky mountains of unstructured data. To get a sense of what is possible we can take a look at a leading transparency mashup called MAPLight.org.

The MAP in MAPLight.org stands for "money and politics," and the site's mission is to illuminate the connection between the two.[97] Founded by computer expert

[93] Robert D. Hof, *Mix, Match, and Mutate*, BusinessWeek, Jul. 25, 2005, at 72, *available at* http://www.businessweek.com/magazine/content/05_30/b3944108_mz063.htm.

[94] For a catalog of the many mashups available on the Internet, *see* ProgrammableWeb – Mashups, APIs, and the Web as Platform, http://www.programmableweb.com (last visited Oct. 10, 2007).

[95] *See* Book Burro, http://bookburro.org (last visited March 5, 2008).

[96] *See, e.g.*, Trendio, About Trendio, http://www.trendio.com/blogs/jensen/?page_id=5 (last visited March 7, 2008) ("Trendio is the first current events stock exchange.").

[97] *See* MAPLight.org, About MAPLight.org, http://maplight.org/about (last visited Oct. 10, 2007).

Dan Newman,[98] the site mashes together congressional voting data from GovTrack.us and campaign finance information from OpenSecrets.org, in addition to information from other sources.[99] The result is a searchable database that highlights the connections between campaign contributions and how members of Congress vote.

Using the MAPLight database, users can look up a particular bill and see the interest groups, as well as the individuals and corporations, who support and oppose it. For example, one can look up H.R. 5252, the Communications Opportunity, Promotion, and Enhancement (COPE) Act of 2006 in the 109th Congress.[100] The groups listed supporting the bill included "Electronics manufacturing & services" and "Farm organizations and cooperatives," while groups listed opposing it included "Consumer groups" and "Online computer services." Drilling down, the Consumer Electronics Association and the National Grange are listed as supporters, while Common Cause and Google are listed in the opponent's column. By clicking on a "Votes" tab, the viewer is shown the last vote on the bill, including how much money the supporting and opposing groups contributed to the campaigns of legislators voting for and against the bill.[101] (See Figure 1.) For H.R. 5252, MAPLight shows that groups who supported the bill gave twice as much money to legislators who voted for the bill than they gave to those who opposed it.[102] Counterintuitively, MAPLight also shows that groups opposing the bill gave slightly more money to legislators voting for the bill.[103]

[98] *See generally* Joan Hamilton, *Politics Watchdog Follows the Money Online*, The Nation, June 11, 2007, http://www.thenation.com/doc/20070625/hamilton (last visited March 7, 2008) (profiling Dan Newman and his inspiration to create MAPLight.org).

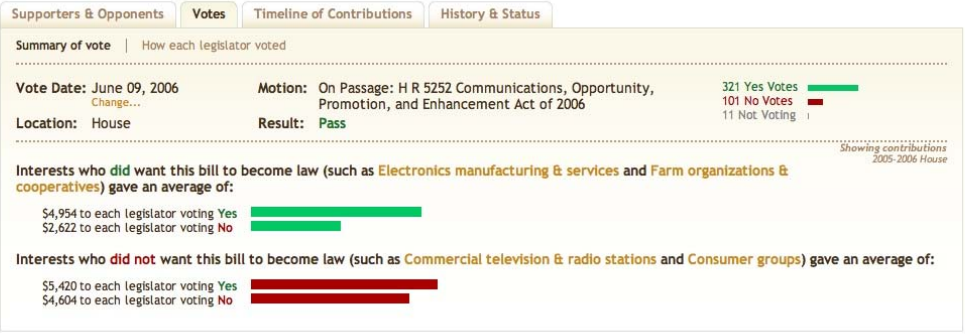
[99] MAPLight.org, Data Sources, http://www.maplight.org/sources (last visited Oct 10, 2007).

[100] MAPLight.org, Supporters and Opponents of H.R. 5252, 109th Cong. (2006), http://www.maplight.org/map/us/bill/40340/default (last visited Oct. 10, 2007).

[101] MAPLight.org, Votes on H.R. 5252, 109th Cong. (2006), http://www.maplight.org/map/us/bill/40340/default/votes/vote-292477 (last visited Oct. 10, 2007).

[102] *Id.* (showing that groups who supported the bill gave an average of $7,192 to each legislator voting for the bill and an average of $3,742 to each legislator voting against it).

[103] *Id.* (showing that groups who opposed the bill gave an average of $5,420 to each legislator voting for the bill and an average of $4,604 to each legislator voting against it).



Vote Date: June 09, 2006

Location: House

Motion: On Passage: H R 5252 Communications, Opportunity, Promotion, and Enhancement Act of 2006

Result: Pass

321 Yes Votes
101 No Votes
11 Not Voting

Showing contributions 2005-2006 House

Interests who did want this bill to become law (such as Electronics manufacturing & services and Farm organizations & cooperatives) gave an average of:

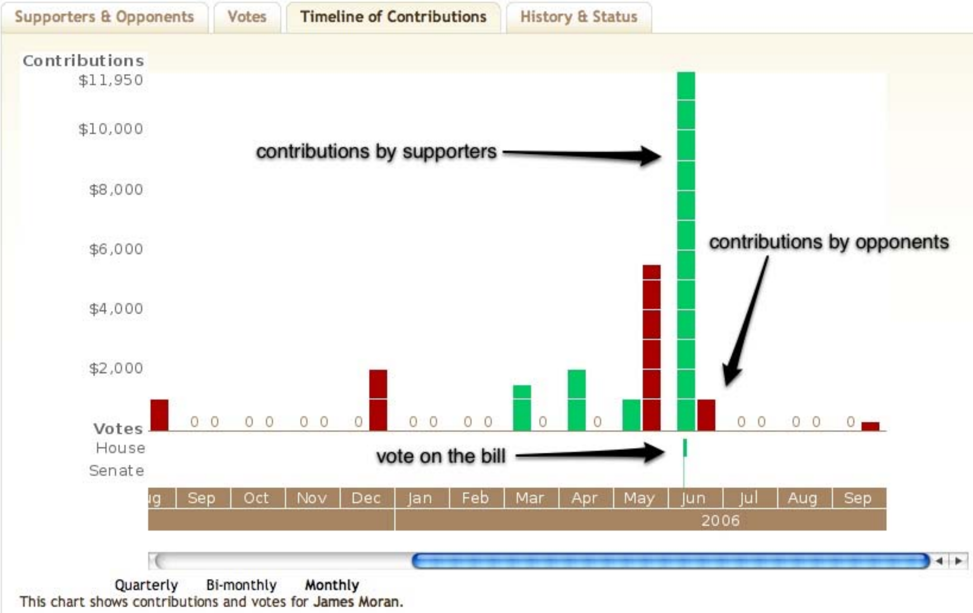
$4,954 to each legislator voting Yes
$2,622 to each legislator voting No

Interests who did not want this bill to become law (such as Commercial television & radio stations and Consumer groups) gave an average of:

$5,420 to each legislator voting Yes
$4,604 to each legislator voting No

***Figure 1** – MAPLight.org page for the June 9, 2006 vote on H.R. 5252. For each vote on a bill, MAPLight.org shows the average amount of money contributed to legislators voting for and against the bill arranged by contributors' positions on the bill.*

MAPLight also allows users to look up individual members of Congress, in order to see how they voted on a particular bill and to see how much money they received from groups supporting and opposing the bill. More to the point, MAPLight offers a timeline of contributions and votes. Figure 2 displays a time line of contributions to Rep. Jim Moran (D–Va.) and votes on H.R. 5252.[104] The bars represent contributions by groups and individuals supporting the bill for a particular time period (in this case monthly), while flags represent votes taken on the bill.

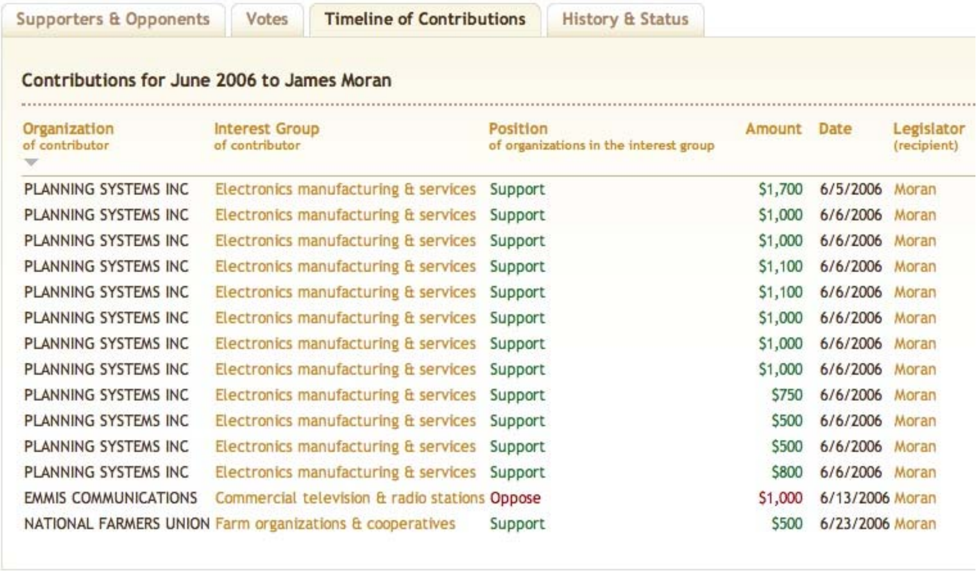


This chart shows contributions and votes for James Moran.

***Figure 2** – MAPLight.org timeline of contributions to Rep. Jim Moran before and after the June 9, 2006 vote on H.R. 5252. Bars representing contributions are color-coded green to indicate contributions by supporters of the bill, and red to indicate contributions by opponents.*

[104] MAPLight.org, Timeline of Contributions to James Moran, http://www.maplight.org/map/us/bill/40340/default/timeline/380?scale=1 (last visited Oct. 10, 2007) (charting a timeline of contributions to Congressperson Moran against Moran's vote on H.R. 5252).

In Figure 2, the flag's color is green, to indicate Moran voted yes on the bill. Hovering the mouse cursor over the flag would display that the vote was taken on June 9th of 2006. Clicking on the bars for June leads to a list of contributions, shown in Figure 3. The webpage shows that Rep. Moran received $11,450 in contributions from either Planning Systems, Inc. or persons affiliated with that company, three days before the vote on the bill.[105]



Contributions for June 2006 to James Moran

Organization of contributor	Interest Group of contributor	Position of organizations in the interest group	Amount	Date	Legislator (recipient)
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$1,700	6/5/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$1,000	6/6/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$1,000	6/6/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$1,100	6/6/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$1,100	6/6/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$1,000	6/6/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$1,000	6/6/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$1,000	6/6/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$750	6/6/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$500	6/6/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$500	6/6/2006	Moran
PLANNING SYSTEMS INC	Electronics manufacturing & services	Support	$800	6/6/2006	Moran
EMMIS COMMUNICATIONS	Commercial television & radio stations	Oppose	$1,000	6/13/2006	Moran
NATIONAL FARMERS UNION	Farm organizations & cooperatives	Support	$500	6/23/2006	Moran

***Figure 3** – MAPLight.org detail view for contributions made to Rep. Jim Moran before and after the June 9, 2006 vote on H.R. 5252.*

This is not to suggest that anything improper occurred with regards to this particular bill, but rather is meant simply to highlight the power of an Internet application that can tap into congressional voting data and campaign finance data and mash them into a tool that citizens can use to illuminate potential connections. This is a new window into congressional actions that legislators did not previously need to consider. Such a mashup would not be possible without the structured data that government often fails to provide and that is being made accessible by hacks such as GovTrack.us and OpenSecrets.org.

Another mashup aimed at increasing government transparency is OpenCongress.org.[106] Among other things, this site takes bill and vote data from GovTrack.us and mashes it with data feeds from blogs and mainstream news sources; so that one can pull up a page for a bill or a legislator and see news stories and blog posts that mention the bill and/or legislator.

[105] MAPLight.org, Timeline of Contributions to James Moran for June 2006, http://www.maplight.org/map/us/bill/40340/default/timeline/380-2006-6-6 (last visited Oct. 10, 2007).

[106] *See* Sunlight Foundation & Participatory Politics Foundation, OpenCongress, http://www.opencongress.org (last visited March7, 2008).

Current mashups, such as MAPlight.org and Opencongress.org, demonstrate the potential for future mashup applications. As long as the information is available in an open and structured format, developers will be able to mash together data sets in unpredictable ways. Such mashups can highlight patterns that would be otherwise imperceptible in the source data.

C. Crowdsourcing

If government data is successfully opened to public scrutiny online–either by official publication or by hacks–seemingly impenetrable mountains of data will be made available. Mashups can help ease the information overload by highlighting the most interesting connections among data sets, but human judgment is still necessary to determine the most relevant facts. Crowdsourcing[107] presents the key to sifting through the data made available by official disclosures, hacks, and mashups.

In early 2007, the scandal concerning eight U.S. Attorneys fired for allegedly political reasons, which some would say ultimately led to Attorney General Alberto Gonzalez's resignation, was beginning to simmer.[108] The Senate Judiciary Committee, which was investigating the matter, was locked in a confrontation with the Department of Justice (DOJ) and the White House over the release of e-mails and other documents related to the firings.[109]

The Justice Department finally relented and at 8:30 p.m. on Monday, March 19th, it delivered to Congress 3,000 pages of e-mails, memos, and notes related to the firings.[110] The Justice Department delivered only one set of the documents and they were not organized in any particular fashion.[111] Immediately, congressional staffers began scanning the paper documents, putting the pages on the Judiciary Committee's website.[112] They completed this task around 1 a.m. that night.[113]

[107] The term "crowdsourcing" was coined by Jeff Howe in a *Wired* magazine article. Jeff Howe, *The Rise of Crowdsourcing*, Wired, June 2006, at 176, *available at* http://www.wired.com/wired/archive/14.06/crowds.html.

[108] *See* David Johnston, *Dismissed U.S. Attorneys Received Strong Evaluations*, N.Y. Times, Feb. 25, 2007, at A19.

[109] *See* Sheryl Gay Stolberg, *White House Delays Action in Inquiry on Attorneys*, N.Y. Times, Mar. 17, 2007, at A12.

[110] *See* Elizabeth Williamson, *Just 3,000 Pages Until Bedtime*, Wash. Post, Mar. 21, 2007, at A13.

[111] *Id.*

[112] *Id.*

[113] *Id.*

A cynic might say that this was a "document dump"[114] meant to obfuscate and lessen the impact of the disclosure. Reporters also stayed up late that night waiting on Capitol Hill and at their offices for the documents they would then have to plow through. According to one House staffer, congressional aides "felt that the late, paper-only release was done more to thwart the media than the committee."[115]

The media, however, was not thwarted; and the next morning's papers included articles detailing some of the e-mails among the documents.[116] The first with relevant analysis of the documents, however, were blogs. Most notable was TPMMuckraker.com, a site that had been following the scandal since its very beginning.[117] TPMMuckraker.com calls itself "a news blog dedicated to chronicling, explaining and reporting on public corruption, political scandal and abuses of the public trust of all sorts."[118] It was started by prominent liberal blogger Josh Marshall of TalkingPointsMemo.com.

On the evening of March 19th, Marshall and co-blogger Paul Kiel were readying themselves for the disclosure avalanche. On the blog that night, Kiel wrote, "Josh and I were just discussing how in the world we are ever going to make our way through 3,000 pages when it hit us: we don't have to. Our readers can help."[119]

In a 12:51 a.m. blog posting titled, "TPM Needs YOU to Comb Through Thousands of Pages," they asked their readers to take small chunks of the documents being made public on the Judiciary Committee's website, read them, and post a comment noting what pages they read and what they found.[120] Kiel wrote:

> And to make it efficient and comprehensible, we'll have a system.

[114] Wikipedia, Document Dump, http://en.wikipedia.org/wiki/Document_dump (last visited Oct. 10, 2007) (defining document dump as "responding to an adversary's request for information by presenting the adversary with a large quantity of data that is transferred in a manner that indicates unfriendliness, hostility, or a legal conflict between the transmitter and the receiver of the information").

[115] Elizabeth Williamson, *Just 3,000 Pages Until Bedtime*, Wash. Post, Mar. 21, 2007, at A13.

[116] *See, e.g.*, David Johnston et al., *New E-Mail Gives Dismissal Detail*, N.Y. Times, Mar. 20, 2007, at A1 (reporting on the contents of several of the disclosed e-mails).

[117] *See, e.g.*, Justin Rood, *Questions, Concerns Swirl around Politics of Prosecutor's Forced Exit*, TPM Muckraker, Jan. 13, 2007, 08:38 EST, http://www.tpmmuckraker.com/archives/002335.php (reporting on the controversy in January of 2007).

[118] About TPM Muckraker, http://www.tpmmuckraker.com/about.php (last visited Oct. 10, 2007).

[119] Paul Kiel, *TPM Needs YOU to Comb Through Thousands of Pages*, TPM Muckraker, Mar. 20, 2007, 00:51 EDT, http://www.tpmmuckraker.com/archives/002809.php.

[120] *Id.*

> As you can see on the House Judiciary Committee's website, they’ve begun reproducing 50-page pdf files of the documents with a simple numbering system, 3-19-2007 DOJ-Released Documents 1-1, then 1-2, then 1-3, etc. So pick a pdf, any pdf and give it a look. If you find something interesting (or damning), then tell us about it in the comment thread below.
>
> Please begin your comment with the pdf number and please provide the page number of the pdf.
>
> . . . .
>
> If you want to be a trailblazer and read through a virgin pdf, then you should be able to see which pdfs haven’t been looked at by scrolling through the comment thread. Have at it![121]

The site’s readers responded immediately and began logging in their contributions. Comments ranged from descriptions of the documents, like, “1-6 Page 21. Margaret Chiara struggles to figure out the real reason for her dismissal in a Feb. 1 e-mail. Sad,”[122] to quotations of the most relevant e-mail messages.[123] By the next morning, almost all of the documents had been read at least once. Kiel had gone to bed after writing the blog post asking for help, and the next morning when he began work again at 7:30, the new user-created resource was waiting for him.[124]

“We have readers on the west coast and readers in Europe and some that are up all night I guess. . . . We had a couple of hundred comments by morning,” Kiel says.[125] Kiel and an intern began reading all the reader comments and whenever they would come across an interesting finding, Kiel would pull the original document and write a story for the blog. He says, “You can see the day after the document dump I published five or six stories on what the [dumped] e-mails were saying, and I wouldn’t have been able to do that if I was spending all my time reading through all the e-mails.”[126]

Among the various reader discoveries was an 18-day gap in the emails that were included in the document dump.[127] This gap coincided with the time right before the U.S.

[121] *Id.*

[122] Comment of TK to Paul Kiel, *TPM Needs YOU to Comb Through Thousands of Pages*, TPM Muckraker, Mar. 20, 2007, 01:14 EDT, http://www.tpmmuckraker.com/archives/002809.php.

[123] *See* Comment of JPV to Paul Kiel, *TPM Needs YOU to Comb Through Thousands of Pages*, TPM Muckraker, Mar. 20, 2007, 01:08 EDT, http://www.tpmmuckraker.com/archives/002809.php.

[124] Telephone Interview with Paul Kiel, Editor, TPMMuckraker.com (Aug. 10, 2007).

[125] *Id.*

[126] *Id.*

[127] Josh Marshall, *Untitled*, Talking Points Memo, Mar. 21, 2007, 03:05 EDT,

attorney firings. "Someone actually pointed that out in the comments," Kiel says, "and that's something I don't think I would have noticed otherwise."[128]

This is an example of what has become known as "crowdsourcing," or, in academic circles, "peer production."[129] The idea is to allow a large group of persons to create, by making small individual contributions, a good that would traditionally have been produced by a single individual or an organization.[130] Usually, the goods in question are cultural or informational products.[131] Wikipedia, the online community-written encyclopedia, is the most often cited example of successful crowdsourcing.[132] Thousands of volunteers labor for no monetary compensation to write basic reference articles for every topic under sun. The result is an encyclopedia that is much more extensive than anything a traditional organization with a limited number of writers and editors could produce.[133]

This sort of collaboration is possible because the Internet has dramatically reduced the transaction cost of interaction between individuals.[134] Persons engaged in collaborative projects such as Wikipedia are often motivated by incentives other than cash compensation, including gaining a positive reputation within a community, and the intrinsic joy of creation and participation.[135] Additionally, participation in non-compensated collaborative projects will often result in lucrative ancillary work.[136] For

http://www.talkingpointsmemo.com/archives/013171.php.

[128] Telephone Interview with Paul Kiel, Editor, TPMMuckraker.com (Aug. 10, 2007).

[129] The term 'peer production' was coined by Prof. Yochai Benkler in his seminal paper describing this new form of production. Yochai Benkler, *Coase's Penguin, Or, Linux and the Nature of the Firm*, 112 Yale L.J. 369 (2002).

[130] Don Tapscott & Anthony D. Williams, Wikinomics: How Mass Collaboration Changes Everything 67 (2006).

[131] *Id.* at 70 ("Peering works best when at least three conditions are present: 1) The object of production is information or culture, which keeps the cost of participation low for contributors . . . ."). Benkler's study of the peer production model is focused on the production of "information and culture." Benkler, Coase's Penguin, *supra* note 129, at 375-78.

[132] *See* Benkler, *supra* note 129, at 386-87, 440-43; Tapscott & Williams, *supra* note 130, at 65-67.

[133] Tapscott & Williams, *supra* note 130, at 71-72.

[134] Martin Kenney & James Curry, *Beyond Transaction Costs: E-commerce and the Power of the Internet Dataspace* 8 (Berkeley E-conomy Project Working Paper No. 18, 2000), *available at* http://e-conomy.berkeley.edu/publications/wp/internet_and_geography.pdf.

[135] Benkler, *supra* note 129, at 423-43.

[136] Benkler, *supra* note 129, at 372, 424-25 & 433; Tapscott & Williams, *supra* note 130, at 83-85.

example, IBM eliminated its traditional proprietary operating system and server development and instead assigned its engineers to contribute to the freely distributed Linux operating system and Apache server platform.[137] The company's revenue comes from the selling hardware that runs open source software, as well as expert support services.[138]

Another celebrated instance of crowdsourcing is the story of Goldcorp. In 1999, the small Toronto-based mining company was on the verge of bankruptcy, having seemingly exhausted its once lucrative gold mine. The company's new CEO, Rob McEwen, was frustrated that the company's in-house geologists could not find the location of gold on the mine nor estimate its value. So he did something that was unthinkable in the mineral extraction business: he published all of the company's proprietary geological data about the mine on the web. The company offered over half a million dollars in prize money to those who could find the gold. Word about the "Goldcorp Challenge" spread quickly and soon hundreds of amateur and professional geologists, academics and retirees, were sifting through the data in ways the in-house geologists could not match. Other areas of expertise were also brought to bear on the problem as mathematicians, physicists, and others tried their hand.

Participants in the challenge identified 55 target areas of the mine that were previously untried by the company. Eighty percent of those newly identified targets yielded substantial amounts of gold and resulted in the company's turnaround.

> McEwen estimates the collaborative process shaved two to three years off their exploration time.
>
> Today Goldcorp is reaping the fruits of its open source approach to exploration. Not only did the contest yield copious quantities of gold, it catapulted his underperforming $100 million company into a $9 billion juggernaut while transforming a backward mining site in Northern Ontario into one of the most innovative and profitable properties in the industry.[139]

The Goldcorp Challenge is much like TPM Muckraker's challenge to its readers. Both sought to sift large quantities of data in order to find valuable nuggets hidden within. Both departed from traditional approaches that rely on managed professionals such as geologists and journalists respectively. In both cases many hands made for light work.

What is different about the TPM Muckraker story is the source of the data. In that case, it was government information that was made public. The Justice Department did not release its emails and memos in electronic format as it could have. Instead they released the documents as paper printouts, which cannot easily be shared electronically. Even after congressional staffers scanned the documents into electronic files, they were

[137] Tapscott & Williams, *supra* note 130, at 77-83.

[138] *Id.*

[139] *Id.* at 9.

images and not text documents that could be searched.[140] Despite these hurdles, however, enough volunteers working simultaneously on discrete chunks of the data could find and highlight all the relevant portions.

While there were only 1,452 daily newspapers in the United States as of 2005,[141] there are now about 70 million blogs in operation, and about 120,000 new blogs come online each day.[142] The vast majority of these blogs no doubt serve to inform and entertain a small circle of friends and relatives.[143] Nevertheless, tens of thousands aspire to engage in journalism and some have been successful.[144] What this affords is a massive pool of ready and willing citizen journalists the likes of which traditional media has never assembled. This strength in numbers can allow the new technologies of transparency to be put to fruitful use despite the quantity of data available.

In his seminal essay, *The Cathedral and the Bazaar*, Eric S. Raymond contrasts the open source method of software development–in essence peer production or crowdsourcing–to the traditional hierarchical model.[145] In the former, a large number of developers contribute simultaneously to the formulation and testing of software code, while central organization and a small number of developers typify the latter. He explains that one of the key differences is the number of eyes sifting through code looking for problems and solutions. He proposes what he calls "Linus' Law": "Given enough eyeballs, all bugs are shallow."[146] Raymond writes,

[140] *See supra* notes 29-31 and accompanying text.

[141] Newspaper Association of America, The Source: Newspapers by the Numbers 22 (2006), http://www.naa.org/thesource/the_source_newspapers_by_the_numbers.pdf.

[142] Scott Gant, We're All Journalists Now: The Transformation of the Press and Reshaping of the Law in the Internet Age 25 (2007); David Sifry, *The State of the Live Web, April 2007*, Sifry's Alerts, Apr. 5, 2007, 02:02 PDT, http://www.sifry.com/alerts/archives/2007/04/the_state_of_th.html.

[143] Gant, supra note 142, at 25-26.

[144] As Scott Gant notes,

> The growing importance of blogging as a source of news and opinion is evident not just from the number of blogs and their readers. It is also evident from polling conducted during January and February 2007, which found 30 percent of respondents view blogging as an important source of news and information (the figure was above 40 percent for those ages 18-29), while more than 55 percent identify it as important to the future of journalism (65 percent of those ages 18-29).

Id. at 26.

[145] Eric S. Raymond, The Cathedral and the Bazaar: Musings on Linux and Open Source By An Accidental Revolutionary 21-22 (rev. 2001), *available at* http://catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/ar01s04.html.

[146] *Id.* at 30.

> In Linus' Law, I think, lies the core difference underlying the cathedral-builder and bazaar styles. In the cathedral-builder view of programming, bugs and development problems are tricky, insidious, deep phenomena. It may take months of scrutiny by a dedicated few to develop confidence that you've winkled them all out. Thus the long release intervals, and the inevitable disappointment when long-awaited releases are not perfect.
>
> In the bazaar view, on the other hand, you assume that bugs are generally shallow phenomena–or, at least, that they turn shallow pretty quick when exposed to a thousand eager co-developers pounding on every single new release. Accordingly you release often in order to get more corrections, and as a beneficial side effect you have less to lose if an occasional botch gets out the door.[147]

Given enough eyeballs, corruption and waste are similarly shallow problems. In the cathedral-builder view of journalism, corruption is hidden from a relatively small number of practitioners by the inaccessibility of government data and the sheer volume of it. In the bazaar view, a vast number of eyes, aided by hacks and mashups, make the amount of data less daunting. The number of eyeballs comes not just from bloggers aiming to do journalism (although they are likely the most dedicated) but also from average citizens contributing to interactive sites.

These interactive websites have begun to leverage what James Surowiecki calls the "wisdom of the crowds" to shed light on government data.[148] For example, WahingtonWatch.com gathers data on bills pending before Congress and mashes them with Congressional Budget Office estimates on the cost of each bill in order to present average cost of bills per family or individual. Aside from presenting this information, the site allows users to contribute by registering their support for or opposition to bills and by posting comments about bills. More importantly, the site is also a wiki for pending legislation. Each bill's page contains a detailed summary of the bill, the bill's status, and points in favor and against, all of which can be edited or added to by anyone. Congresspedia.com is a similar community-written wiki that also includes biographical pages for members of Congress.

These sites are community-created collection buckets for the interesting and essential bits of information that surface from the gigabytes of unsorted government data. While permanent and systematic, such sites are similar to the ad hoc comments thread begun by Marshall and Kiel at TPM Muckraker. They allow users to contribute as much or as little of their time as they would like. Their work is made available to the public, including bloggers and journalists who are "higher up the food chain," and who will

[147] *Id.* at 31.

[148] In his book of the same name, Surowiecki puts forth the thesis that large groups of people are usually as good as, or better than, experts at solving problems and predicting the future. James Surowiecki, The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations (2004).

reward useful information with attention.[149] Also, because users of such community sites will self-select to contribute to subjects about which they are well informed, or in which they have a personal interest or stake, they will keep each other honest, expose all sides of an issue, and hopefully improve the quality of the peer-produced outcome.[150]

The community dissection of the DOJ data dump at TPM Muckraker's behest is an example of success for what can be called the "quick-sifting" crowdsourcing model. On the other hand, transparency-focused community sites are still in their infancy. They have not produced extensive analyses of every, or even most, bills pending in Congress. That said, Wikipedia demonstrates that the "resource-building" model can be effective.

III. RECOMMENDATIONS

As we have seen, Internet technologies have the potential to greatly improve transparency by making government data more accessible and by fostering communities that can identify and highlight relevant information. For this to work, however, certain foundational elements need to be in place. Ideally, government would provide the necessary informational building blocks. After all, it is the source of the data and it could ensure its completeness and accuracy.[151] Government has the power to enact reforms to

[149] Telephone Interview with Paul Kiel, Editor, TPMMuckraker.com (Aug. 10, 2007) (explaining that the TPM Muckraker was monitored by more mainstream journalists covering the story). As noted above, a contributor to the TPM Muckraker comments thread was first to spot the 18-day gap in e-mails handed over by DOJ to Congress. *See supra* note 128 and accompanying text. Subsequently, several news stories reported the fact. *E.g.* Eric Lipton and David Johnston, *Democrats See a 'Document Gap' in Dismissals*, N.Y. Times, Mar. 22, 2007, at A1 ("From Nov. 16 to Dec. 7, there are only a handful of e-mail messages, a fact that Talking Points Memo, a Web site that has been following the furor with microscopic attention, pointed out Wednesday morning.")

[150] Wikipedia, for example, allows a user to track recent changes to all articles or to particular articles. This ensures that interested users can easily spot mistakes and vandalism. One study by IBM researchers found that "[t]he site is subject to frequent vandalism and inaccuracy, just as skeptics might suspect—but the active Wikipedia community rapidly and effectively repairs most damage. Indeed, one type of malicious edit we examined is typically repaired within two minutes." Viégas et al., *Studying Cooperation and Conflict Between Authors with History Flow Visualizations*, 6 CHI Letters 575, 575-76 (2004), *available at* http://alumni.media.mit.edu/~fviegas/papers/history_flow.pdf.

[151] As the Open House Report notes,

> The only freely available source for downloading structured legislative data is created and maintained by GovTrack.us, a private, independent effort. GovTrack's database is the source for the information behind other public Web sites, such as OpenCongress.org, and as a result any errors in the original database have a wide impact. Common errors include delayed bill records, outdated cosponsor lists and incomplete committee membership listings. The errors, gaps and delays stem from the automated way in which the independent databases are reconstructed from the scattered, unstructured information

make the data it produces easily open to the public. If it does not do this, however, the private sector should fill the breach.

A. Defining the Foundation

The first building block of a foundation on which Internet technologies can help improve transparency is the idea that, to the greatest extent feasible, government data should be made public. As we have seen, however, data can be made *technically* available to the public, but generally out of its reach. Data should instead be made *meaningfully* publicly available and in today's day and age this means it should be made available online. Government, however, continues to lag.

For example, the Freedom of Information Act (FOIA) recognized that short of a few exceptions (including for national security and personal privacy) all government data should be available to the public.[152] Of course, under the Act a citizen must file a request for information, and a response can take months or years.[153] The 1996 E-FOIA amendments to the Act were aimed at giving meaning to the notion of publicly available information. The reform required government agencies to publish on their websites the most often requested documents.[154] Not only would doing so increase transparency, but by putting online documents that would likely be requested again, agencies would save resources spent on complying with FOIA requests. Also, while FOIA already mandated that opinions and orders, statements of policy, and staff manuals be made available for public inspection, the E-FOIA Amendments added the requirement that they be available online.[155]

The results, however, have been poor. A 2007 survey of 149 agency websites by the National Security Archive at George Washington University "found massive non-compliance with E-FOIA."[156] Only a fifth of the agencies reviewed made available on

> that is available now. An authoritative structured database directly from Congress would provide a current, complete, accurate and reliable basis for these applications.

Open House Report, *supra* note 19, at 13.

[152] Freedom of Information Act, 5 U.S.C. § 552 (2006).

[153] *See* Nat'l Sec. Archive, *A FOIA Request Celebrates Its 17th Birthday: A Report on Federal Agency FOIA Backlog: Oldest Unanswered Freedom of Information Act Requests Were Filed in 1989* (2006), *available at* http://www.gwu.edu/~nsarchiv/NSAEBB/NSAEBB182/executive_summary.pdf.

[154] 5 U.S.C. § 552(a)(2)(D) (2006).

[155] 5 U.S.C. § 552(a)(2).

[156] Nat'l Sec. Archive, *File Not Found: 10 Years After E-FOIA, Most Federal Agencies Are Delinquent* 1 (2007), http://www.gwu.edu/~nsarchiv/NSAEBB/NSAEBB216/e-foia_audit_report.pdf.

their websites all the data required by FOIA.[157] According to the report, 41 percent of agencies had not posted their most requested documents as FOIA mandates.[158]

There is no excuse for government's failure not to put data online. Almost all data today is created electronically using word processors and other computer applications. Because documents enter the world digitally, the initial step of online publication (i.e., digital formatting) is complete. The next steps, which include designing and implementing useful websites to host the data, should also come at minimal cost since most agencies already have online presences. The rest of the world has come to understand that electronic dissemination of data presents efficiencies and savings over paper, and government should be no different.

The second building block needed for a solid foundation of government data is the idea that information should not just be made available online, but that online resources must also be useful. This means putting data online in structured, open, and searchable formats.

Structured, as we have seen, means that the data is presented in a machine-readable format that makes it easy for individuals to subscribe to discrete data feeds, and for others to use the data in their own creations–that is, as the source data for a community site such as WashingtonWatch.com or mashups like MAPLight.org.

Open means that the digital formats chosen should be non-proprietary and widely accepted. Open formats are often created and maintained by independent standards organizations and are free of copyright restrictions on their use. For example, MP3 is an open audio file format, while RealMedia and Apple QuickTime are proprietary.

There are several reasons to prefer open formats. One is that proprietary formats can often only be opened and viewed reliably with proprietary software. For example, opening Word, Excel, or PowerPoint documents, requires Microsoft Office (which retails between $100 and $300).[159] One could use a free alternative, such as OpenOffice, to manipulate these document types, but compatibility is not perfect.[160] This is a result of the closed nature of the Microsoft formats, which must be reverse engineered. Open formats, on the other hand, are generally international standards in the public domain that can be freely used by anyone.

Another reason to prefer open formats is that if the owner of a proprietary format chooses not to develop a reader or player for their format for a particular computer operating system, users of that system will not be able to access information encoded in that format. For example, the popular iTunes Music Store sells music in a proprietary audio file called protected AAC. If you buy a song from iTunes that is encoded in this format, it will play on Windows and Macintosh computers because Apple has developed

[157] *Id.* at 7.

[158] *Id.* at 1.

[159] Microsoft, however, does make available a free viewer application for its Office suite documents, but not for all computing platforms, including Linux.

[160] Edward Mendelson, *Review of OpenOffice.org 2.3*, PC Magazine, Sept. 28, 2007, http://www.pcmag.com/article2/0,1895,2190711,00.asp.

the iTunes player for both those platforms. It will not play on computers running Linux, however, because Apple has chosen not to develop iTunes for that operating system. Similarly, protected AACs will only play on iPods because Apple has also chosen not to license the format to other digital music player manufacturers. That is a perfectly legitimate choice for a company to make about its products, but government information should be made available to the largest number of persons possible and at the lowest cost.

Finally, searchable means what it sounds like. The data made available should be full-text searchable to the greatest extent feasible.[161] This should not be an issue if the data is kept in a digital text format once it is created. To the extent that paper documents are scanned into digital files, optical character recognition should be applied to produce searchable text.

B. Laying the Foundation de Jure

A foundation that allows Internet technologies to be leveraged to increase transparency requires government data to be made available online in a structured, open, and searchable format. The most obvious route to this goal is legislation that mandates online disclosure. Any such legislation, however, must take care to ensure that it lays all parts of the foundation.

An example of such pathbreaking legislation is the recently enacted Federal Funding Accountability and Transparency Act of 2006, which requires the online disclosure of all organizations receiving federal funds.[162] It is targeted at legislative earmarks, which Congress uses to direct federal money to specific persons or projects.[163] The Act mandates that OMB establish a searchable website that catalogs each funding award along with relevant information, including the Congressional district in which the money is spent.[164] Its drafters astutely defined the term "searchable website" in the Act and included in its meaning the ability for users to search awards by a number of useful fields.[165] This requirement means that all text will have to be machine-readable and fully searchable. The definition of "searchable website" also requires that the data produced by

[161] *See* Cary Coglianese, *E-Rulemaking: Information Technology and Regulatory Policy* 15-18 (Harvard University John F. Kennedy School of Government Center for Business and Government Regulatory Policy Program, Report No. RPP-05, 2004), *available at* http://www.hks.harvard.edu/m-rcbg/rpp/erulemaking/papers_reports/E_Rulemaking_Report2004.pdf.

[162] Federal Funding Accountability and Transparency Act of 2006, Pub. L. No. 109-282, 120 Stat. 1186 (2006).

[163] Press Release, Senator Barack Obama, Senate Passes Coburn-Obama Bill to Create Internet Database of Federal Spending (Sept. 8, 2006), *available at* http://obama.senate.gov/press/060908-senate_passes_c/.

[164] Federal Funding Accountability and Transparency Act § 2(b)(1).

[165] *Id.* §§ 2(a)(3)(A)-(C).

searches be downloadable.[166] Additionally, the Act requires that the spending data be contained within the centralized site; mere links to data on other government websites will not suffice to comply with the Act.[167] Both of these requirements imply, although they do not make it explicit, that the website's data must be offered in a structured format. How the site is implemented by OMB will ultimately determine how useful it will be to individuals and how easily third parties will be able to re-use and remix the data available there.[168]

In contrast, the E-Government Act of 2002 sought to use "information technology to increase access, accountability, and transparency" at regulatory agencies.[169] To that end it mandated that agencies make available their regulatory dockets online.[170] As noted in Part I.B, *supra*, public access to the regulatory dockets of federal agencies leaves much to be desired. However, agencies are in likely compliance with the Act because it only requires that docket data be made "publicly available online to the extent practicable as determined by the agency in consultation with the Director [of OMB.]"[171] First, the term "to the extent practicable" is arguably an exception that swallows the rule. More to the point, however, is the fact that a requirement that data be "available online," as we have seen, is not the same as easily accessible, searchable, or available in a useful format.

Legislation aiming to make data usefully available online must be specific about what it requires. When it comes to online disclosure, more than a general statement of policy may be necessary. For example, H.R. 170 currently pending before the House would require the financial disclosure statements currently not available online to be made available "on the Internet in a format that is searchable and sortable."[172] That phrasing implies a structured format and it is used throughout the bill. The bill would, for example, amend the section of the U.S. Code that deals with the FEC's obligation to post campaign finance filings online "by inserting 'in a format that is searchable and sortable' after 'Internet.'"[173] The obvious intention is to make the data available already available online accessible in a structured format, which is a great improvement to disclosure requirements that make only passing mention of the Internet. However, this type of language might be improved by more specificity, such as "searchable and sortable and available for download in a structured and open format, such as XML." Such a construction would not limit the choices of a developer, but would give clearer guidance

[166] *Id.* § 2(a)(3)(D).

[167] *Id.* § 2(c)(2).

[168] Welcome to USASpending.gov, http://www.usaspending.gov/ (last visited Mar. 4, 2008).

[169] E-Government Act of 2002, Pub. L. No. 107-347, § 206, 116 Stat. 2899, 2915-16 (2002).

[170] *Id.* § 206(d).

[171] *Id.*

[172] Sunlight Act of 2007, H.R. 170, 110th Cong. § 3 (2007).

[173] *Id.* § 6.

about what is expected.

One argument against requiring government agencies to make data available online in useful formats is that, as we have seen, the market is already providing these information goods. That is, third parties like GovTrack.us are successfully hacking the data and providing it to the public in useful formats, so why should government take on this role?

There are three main reasons why dissemination of raw data in useful formats is a government role. First, government holds the digital originals of the data and can ensure the integrity and quality of the data made available online. The screen-scraping process used by hackers to gather government data is much like the hand copying of texts by medieval monks; while generally accurate, occasional errors are introduced nonetheless. Copies of documents made available by government, on the other hand, can be completely accurate. Hacked databases such as GovTrack.us are the source of information for mashups, "and as a result any errors in the original database have a wide impact."[174]

Second, while exact figures are difficult to estimate, the marginal cost to the government of presenting its data in a useful format is certainly less than the cost incurred by third parties to devise and maintain clever hacks to siphon otherwise difficult-to-access government data. Finally, not all desirable government data can be hacked and made available by third parties. The major obstacle is that the government has not made some data available online. Online availability is a foundational piece that can only be addressed by government, and to the extent it makes new information available online, as we have just seen, it makes most sense for it to do so in useful formats.

Making government information available online is an activity now being performed by the market that can conceivably be carried out more efficiently by government. However, it is in making raw data available in useful formats that government has a comparative advantage. Rather than simply making data available for third-party use, government might be tempted to incorporate into its offerings tools to sort and analyze data much like mashups or crowdsourced projects have done. To the extent that making such tools available precludes or substitutes raw data, government should show restraint. Rather than offering simply "one best way" to utilize data, government should allow myriad third parties develop innovative tools that utilize the data.

C. Laying the Foundation de Facto

If the government, for whatever reason, fails to make its data publicly available online in useful forms, concerned citizens should fill the breach both in order to increase transparency and to cajole the government to take action. GovTrack.us and OpenSecrets.org are examples of independent third parties addressing a need for congressional and campaign finance data. There are many other parts of the government,

[174] Open House Report, *supra* note 19, at 13.

including the dozens of executive branch agencies and independent regulatory commissions, that third parties can help make more transparent. When citizens take it upon themselves to place government data online in a useful manner, they not only help keep government accountable, they can induce change in government practices.

Carl Malamud is an early pioneer of making government information available online and a good example of how citizens can effect change. Malamud is an economist who has developed software for the Federal Reserve and who started a company to offer regular audio broadcasts over the Internet.[175] Malamud launched Internet Multicasting Service, a non-profit organization that assisted the Securities and Exchange Commission (SEC) with making data available to the public over the Internet.[176]

At that time, the SEC did not provide free access to the corporate filings it collected. Instead, the SEC's database, the Electronic Data Gathering Analysis and Retrieval system (EDGAR), was operated under contract with information wholesaler Mead Data, which provided data feeds to data retailers who in turn sold access to the public.[177] "Under this system, a retail information provider, like Mead Data's own Nexis service, charge[d] about $15 for each S.E.C. document, plus a connection charge of $39 an hour and a printing charge of about $1 a page."[178] As one can imagine, customers were largely restricted to Wall Street Firms.

In January 1994, Malamud began to purchase the SEC's wholesale data and made it available on his website free of charge to anyone.[179] The service included corporate annual reports, 10-K filings, proxy statements, and other data valuable to investors, journalists, and others.[180] Unlike the data provided by commercial retailers, Malamud's website posted data with a 24-hour lag and did not contain any value-added analysis or other services.[181] Later that year in December, Malamud expanded his free offerings by adding large portions of the Patent and Trademark Office's (PTO) patent and trademark database, including full text of all patents and text and images from the trademark database.[182]

[175] John Markoff, *Plan Opens More Data to Public*, N.Y. Times, Oct. 22, 1993, at D1.

[176] *Id.*

[177] *Id.*

[178] *Id.*

[179] *See* Peter H. Lewis, *Internet Users Get Access To S.E.C. Filings Fee-Free*, N.Y. Times, Jan. 17, 1994, at D2 (Malamud paid $78,000 a year for data tapes).

[180] *Id.*

[181] *Id.* In 1995 Malamud began offering same-day services thanks to an anonymous donation. Associated Press, *Same-Day Internet Access to S.E.C. Filings*, N.Y. Times, May 1, 1995, at D5.

[182] John Markoff, *Group to Widen Access To Federal Data Bases*, N.Y. Times, Dec. 23, 1994, at D2.

Malamud, however, believed that it was government's province to provide its data for free to the public, especially since the recently passed Paperwork Reduction Act mandated that agencies make public information available electronically.[183] On August 11, 1995, Malamud announced on his website that it would discontinue its free access to government data on October 1st. As Malamud later recounted,

> Our goal, however, wasn't to be in the database business. Our goal was to have the SEC serve their own data on the Internet. After we built up our user base, I decided it was time to force the issue. That's when the fireworks began. When users visited our EDGAR system in August 1995, they got an interesting message:
>
> This Service Will Terminate in 60 Days
> Click Here For More Information
>
> Click here they did! One of the lessons I've learned from building Internet services is that when people get something for free, they want their money's worth.[184]

The SEC at first resisted.[185] Eventually, however, it relented and the agency took over Malamud's service as the core of an online EDGAR system.[186] The public uproar apparently caught SEC commissioners off guard and they took on the responsibility of making data available before the October deadline.[187] According to Malamud, "The commissioners of the SEC had clearly not been aware of the issue, but there is nothing like pieces in the *Wall Street Journal* and 15,000 messages to the Chairman to raise the profile of an issue."[188]

Malamud had similar plans for his patent and trademark database. In 1998, he wrote to Vice President Al Gore and Commerce Secretary William Daley (who oversaw the PTO), announcing that unless the PTO began offering its databases online, he would

[183] Associated Press, *An Internet Access to S.E.C. Filings to End Oct. 1*, N.Y. Times, Aug. 12, 1995, at D1. *See also* Paperwork Reduction Act, 44 U.S.C. § 3506(d) (2006) (stating agency responsibilities regarding information dissemination).

[184] Carl Malamud, *The Importance of Being EDGAR*, Mappa.Mundi Magazine, Sept. 15, 1999, http://www.mundi.net/cartography/EDGAR.

[185] *Id.*; Associated Press, *supra* note 181.

[186] Malamud, *supra* note 184; s*ee also S.E.C. Seeks to Keep Free Internet Service*, N.Y. Times, Aug. 16, 1995, at D7.

[187] Malamud, *supra* note 184.

[188] *Id.*

create a free robust database online.[189] In the intervening years since Malamud put the database online in 1994, the PTO had not been as accommodating as the SEC, largely because the agency is self-financed by user fees, a large portion of which came from requests for paper copies of patent and trademark information.[190] As the Commissioner for Patents told the *New York Times*, "If he can [put the patent and trademark database online] we'd be out all $20 million we now receive in fees . . . Why would anyone want paper?"[191]

The strategy to overcome government's resistance was a familiar one. "I'm going to buy the trademark data and will build the user base as big as I can in a year,"[192] Malamud said at the time. "At the end of the year, I'll pull the rug out from the users and give them Al Gore's E-mail address."[193] The gambit worked and less than two months later the Clinton administration announced that it would put the full patents database online.[194]

It is interesting to note that Malamud's motivations were not just to increase transparency, but also to unleash the creative forces that today result in mashups. According to a 1998 article in the *New York Times*, "His hope is that by making the entire patent data base available to any college student who has managed to acquire 100 gigabytes of disk storage capacity he will touch off an explosion of creative ways in which to plumb the nation's science and technology storehouse."[195]

More recently, Malamud has tried to cajole congressional leaders to make video of their proceedings online indefinitely. While Congress often streams live video feeds of committee hearings and other proceedings, these videos are not often archived and they simply disappear into the ether once the broadcasted event concludes. The result is that "a lot of significant public business is seen only by the people who happen to be in the room: lobbyists plus a smattering of tourists."[196] Malamud has begun to record congressional video streams and put them online at the Internet Archive and Google

[189] Letter from Carl Malamud, President, Internet Multicasting Service, to Al Gore, Vice President of the United States (Apr. 27, 1998) (on file with author), *available at* http://public.resource.org/letter.html.

[190] John Markoff, *U.S. is Urged to Offer More Data on Line*, N.Y. Times, May 4, 1998, at D6.

[191] *Id.*

[192] *Id.*

[193] *Id.*

[194] John Markoff, *U.S. to Release Patent Data on a World Wide Web Site*, N.Y. Times, Jun. 25, 1998, at D2.

[195] Markoff, *supra* note 190.

[196] James Fallows, *Another Win for Carl Malamud*, TheAtlantic.com, Mar. 9, 2007, http://www.theatlantic.com/doc/200703u/c-span.

Video to demonstrate that there is nothing technical stopping Congress from archiving its videos itself.[197] In March of 2007, Malamud wrote a letter to House Speaker Nancy Pelosi detailing the need for congressional video and how the House could go about implementing an online archive.[198]

With any luck we will see more such independent efforts to bring light to information that the government has kept offline and in the dark. Making data public in useful forms not only increases transparency and allows creative uses of data, but also puts pressure on government to make the data available itself.

Conclusion

To hold government accountable for its actions, citizens must know what those actions are. To that end, they must insist that government act openly and transparently to the greatest extent possible. In the twenty-first century, this entails making its data available online and easy to access. If government data is made available online in useful and flexible formats, citizens will be able to utilize modern Internet tools to shed light on government activities. Such tools include mashups, which highlight hidden connections between different data sets, and crowdsourcing, which makes light work of sifting through mountains of data by focusing thousands of eyes on a particular set of data.

Today, however, the state of government's online offerings is very sad indeed. Some nominally publicly available information is not online at all, and the data that is online is often not in useful formats. Government should be encouraged to release public information online in a structured, open, and searchable manner. To the extent that government does not modernize, however, we should hope that private third parties build unofficial databases and make these available in a useful form to the public.

[197] *See* Internet Archive videos submitted by Carl Malamud, http://www.archive.org/search.php?query=hooptedoodle (last visited Mar. 6, 2008); s*ee also* Kelly McCormack, *Web Pioneer Urges House to Upgrade Video Quality*, The Hill, Mar. 23, 2007, at 7, *available at* http://thehill.com/leading-the-news/web-pioneer-urges-house-to-upgrade-video-quality-2007-03-22.html (describing Malamud's efforts to upload videos to an Internet video archive and Google Video).

[198] Letter from Carl Malamud, Senior Fellow, Center for American Progress, to Nancy Pelosi, Speaker of the House of Representatives of the United States (Mar. 13, 2007) (on file with author), *available at* http://public.resource.org/dear_speaker.html.