GOVERNMENT DATA AND THE INVISIBLE HAND

David Robinson*, Harlan Yu*†, William P. Zeller*†, & Edward W. Felten*†,•

11 YALE J.L. & TECH. 160 (2009)

INTRODUCTION

If President Barack Obama's new administration really wants to embrace the potential of Internet-enabled government transparency, it should follow a counter-intuitive but ultimately compelling strategy: *reduce* the federal role in presenting important government information to citizens. Today, government bodies consider their own Web sites to be a higher priority than technical infrastructures that open up their data for others to use. We argue that this understanding is a mistake. It would be preferable for government to understand providing reusable data, rather than providing Web sites, as the core of its online publishing responsibility.

During the presidential campaign, all three major candidates indicated that they thought the federal government could make better use of the Internet. Barack Obama's platform went the furthest and explicitly endorsed "making government data available online in universally accessible formats."[1] Hillary Clinton, meanwhile, remarked that she wanted to see much more government information online.[2] John McCain's platform called for a new Office of Electronic Government.[3]

But the situation to which these candidates were responding—the wide gap between the exciting uses of Internet technology by private parties, on the one hand, and the government's lagging technical infrastructure, on the other—is not new. A minefield of federal rules and a range of other factors, prevent government Web masters from keeping pace with the ever-growing potential of the Internet.

In order for public data to benefit from the same innovation and dynamism that characterize private parties' use of the Internet, the federal government must reimagine its role as an information

* Center for Information Technology Policy, Princeton University.

† Department of Computer Science, Princeton University.

• Woodrow Wilson School of Public and International Affairs, Princeton University.

[1] Barack Obama and Joe Biden: Technology, http://www.barackobama.com/issues/technology/ (last visited Dec. 2, 2008).

[2] *Meet the Press* (NBC television broadcast Jan. 13, 2008), *available at* http://www.msnbc.msn.com/id/22634967.

[3] JohnMcCain.com: Technology, http://www.johnmccain.com/Informing/Issues/cbcd3a48-4b0e-4864-8be1-d04561c132ea.htm (last visited Dec. 2, 2008).

provider. Rather than struggling, as it currently does, to design sites that meet each end-user need, it should focus on creating a simple, reliable and publicly accessible infrastructure that "exposes" the underlying data. Private actors, either nonprofit or commercial, are better suited to deliver government information to citizens and can constantly create and reshape the tools individuals use to find and leverage public data. The best way to ensure that the government allows private parties to compete on equal terms in the provision of government data is to require that federal Web sites themselves use the same open systems for accessing the underlying data as they make available to the public at large.

Our approach follows the engineering principle of separating data from interaction, which is commonly used in constructing Web sites.[4] Government must provide data, but we argue that Web sites that provide interactive access for the public can best be built by private parties. This approach is especially important given recent advances in interaction, which go far beyond merely offering data for viewing, to providing services such as advanced search, automated content analysis, cross-indexing with other data sources, and data visualization tools. These tools are promising but it is far from obvious how best to combine them to maximize the public value of government data. Given this uncertainty, the best policy is not to hope government will choose the one best way, but to rely on private parties in a vibrant marketplace of engineering ideas to discover what works.

I. FEDERAL INTERNET PRESENCE: THE STATE OF PLAY

The Internet's transformative political potential has been clear to astute nontechnical observers since at least the mid-1990s, but progress toward that transformation has been sporadic at best. In January of 1995, when the Republicans regained a Congressional majority, they launched THOMAS, a Web site that details every bill in Congress.[5] But by 2004, the site was so out of date that seven senators cosponsored a resolution to urge the Library of Congress to modernize it.[6]

The Federal Communications Commission—the agency most closely involved in overseeing digital communications—has

[4] Most sophisticated Web sites use separate software programs for data and interaction, for example storing data in a database such as MySQL, while interacting with the user via a Web server such as Apache. Many government Web sites already use such a separation internally. Government sites that currently separate these functions are already partway to the goal we espouse.

[5] Library of Congress, About THOMAS, http://thomas.loc.gov/home/abt_thom.html (last visited Jan. 3, 2009).

[6] S. Res. 360, 108th Cong. (2004) ("A resolution expressing the sense of the Senate that legislative information shall be publicly available through the Internet.").

a Web site whose basic structure has remained unchanged since 2001.[7] Regular users of the system report that in order to obtain useful information, they must already know the docket number for the proceeding in which they are interested.[8] Materials can be searched by a few criteria such as the date of submission or name of the submitting attorney, but the site does not allow users to search the actual content of comments and filings *even when these filings have been submitted to the agency in a computer-searchable file format*.[9] Even Google, which is severely handicapped by its lack of access to the agency's internal databases, does a significantly better job of identifying relevant information.[10]

Federal Web masters are eager to embrace the Internet's full potential, and in some cases, they have been remarkably successful in the context of their challenging environment. Compared to technologists in the private sector, federal Web masters face a daunting array of additional challenges and requirements. An online compliance checklist for designers of federal Web sites identifies no fewer than twenty-four different regulatory regimes with which all public federal Web sites must comply.[11] Ranging from privacy and usability to FOIA compliance to the demands of the Paperwork Reduction Act and, separately, the Government Paperwork Elimination Act, each of these requirements alone is, considered on its own, a thoughtfully justified federal mandate. Each one reflects the considered judgment of our political process, informed by the understanding of information technology that was available when it was written. But the cumulative effect of these requirements, taken together, is to place federal Web designers in a compliance minefield that

---

[7] *Compare* Wayback Machine Internet Archive for http://www.fcc.gov from September 17, 2001, http://web.archive.org/web/20010917033924/http://www.fcc.gov/, *with* Federal Communications Commission, http://www.fcc.gov/ (last visited Dec. 2, 2008).

[8] *See* Posting of Jerry Brito to Tech. Liberation Front, FCC.gov: The Docket that Doesn't Exist, http://techliberation.com/2007/11/01/fccgov-the-docket-that-doesnt-exist/ (Nov. 1, 2007); Posting of Cynthia Brumfield to IP Democracy, The FCC is the Worst Communicator in Washington, http://www.ipdemocracy.com/archives/002640the_fcc_is_the_worst_communicator_in_washington.php (Sept. 5, 2007, 09:17 EST).

[9] Jerry Brito, *Hack, Mash & Peer: Crowdsourcing Government Transparency*, 9 COLUM. SCI. & TECH. L. REV. 119, 123-25 (2007), *available at* http://www.stlr.org/html/volume9/brito.pdf.

[10] Posting of Jerry Brito to Tech. Liberation Front, FCC.gov: Searching in Vain, http://techliberation.com/2007/10/29/fccgov-searching-in-vain (Oct. 29, 2007).

[11] Web Content Managers Advisory Council, Requirements Checklist for Government Web Managers, http://www.usa.gov/webcontent/reqs_bestpractices/reqs_checklist.shtml (last visited Dec. 2, 2008).

makes it hard for them to avoid breaking the rules—while diverting energy from innovation into compliance. The stultifying compliance climate is an undesirable side effect, not a choice Americans endorsed through our political process.[12] Indeed, there is no guarantee that these requirements interact in such a way as to make total compliance with all of them possible, even in principle.[13]

These problems attend any individual federal Web site; a second layer of challenges can emerge when the federal government seeks to impose coordination or consistency across the remarkably broad range of rulemaking processes and data. This happened with Regulations.gov, a government-wide docket publishing system created in response to the E-government Act of 2002 and launched in 2003. It is used today by "nearly all Departments and Agencies"[14]—in fact, the policy of the Office of Management and Budget (OMB) not only requires its use but also precludes the agencies from using "ancillary and duplicative" docketing and rulemaking systems of their own design.[15] This exclusivity rule, combined with the difficult interagency politics involved in honing system features, have led to a bare-bones approach that leaves out the agency-tailored functionality found in many of the systems it replaced. Concerns about cost-sharing have also led the system to omit even features whose usefulness and desirability is a matter of broad consensus.[16]

Regulations.gov was launched with a limited search engine

---

[12] For example, several different requirements that were developed independently of one another require certain content to be included on homepages. Overall, these rules prevent certain kinds of simple, intuitive interfaces that might in fact be desirable. Our proposal, by reducing the importance of homepages, helps resolve this issue. By making all data available and allowing non-governmental actors to structure interactions around their own aims, information technology professionals can avoid the problem of being mandated to clutter their homepages with boilerplate disclosures.

[13] And compliance is, in any case, a difficult practical challenge. One survey found that only 21% of federal agencies post on the Web all four types of FOIA data required under the 1996 Electronic Freedom of Information Act Amendments. *See* Kristin Adair et al., *File Not Found: Ten Years After E-FOIA, Most Federal Agencies Are Delinquent*, 2007 NAT'L SECURITY ARCHIVE 7, *available at* http://www.gwu.edu/~nsarchiv/NSAEBB/NSAEBB216/index.htm.

[14] Regulations.gov, What Is on This Site, http://www.regulations.gov/search/this_site.jsp (last visited Dec. 2, 2008).

[15] *See* OFFICE OF MGMT. & BUDGET, EXECUTIVE OFFICE OF THE PRESIDENT, EXPANDING E-GOVERNMENT: PARTNERING FOR A RESULTS ORIENTED GOVERNMENT 4 (2004), *available at* http://www.whitehouse.gov/omb/budintegration/expanding_egov12-2004.pdf.

[16] Our discussion of Regulations.gov draws heavily on a recent report by the ABA-chartered Committee on the Status and Future of e-Rulemaking. Cynthia Farina et. al, *Achieving the Potential: The Future of Federal e-Rulemaking*, 2008 SEC. ADMIN. L. & REG. PRAC. AM. BAR ASS'N 1, *available at* http://ceri.law.cornell.edu/erm-comm.php.

and no browsing capability, so that only those who already knew the terms of art used to categorize rulemaking documents were able to use it effectively.[17] Five years later, a re-launched version of the site offered up its limited inventory of computer-readable data directly to the public (in this case, using a single RSS feed) which allowed any interested person or group to create an alternative, enhanced version of the Web site.[18] This has permitted the creation of OpenRegulations.org, which competes with Regulations.gov by offering "paired [sic] down, simple-to-navigate listings of new agency dockets" and a more sensible set of RSS feeds, one for each individual agency.[19]

However, because the engine behind Regulations.gov gathers and integrates only very basic information about the many documents it displays—such as a title, unique identifier, and author name—the decision to share this information with the public can offer only limited benefits. Most of the information relevant to the rulemaking process remains locked away in computer files that are images of printed documents, which cannot be easily reused. A recent ABA-sponsored report concluded that Regulations.gov "continues to reflect an 'insider' perspective"[20] and lacks a comprehensive, full-text search engine over all regulatory data.[21] The same report also emphasized that individual executive branch entities such as the Environmental Protection Agency and the Department of Transportation have been forced to close down their own more advanced systems, which offered deeper insight into docket materials, in order to comply with the prohibition on redundancy.[22] A congressional panel was similarly critical, finding that "[m]any aspects of this initiative are fundamentally flawed, contradict underlying program statutory requirements and have stifled innovation by forcing conformity to an arbitrary government standard."[23]

There are a number of potential ways to improve Regulations.gov. These include changing the funding model so that

---

[17] Ctr. for Democracy & Tech, *A Briefing on Public Policy Issues Affecting Civil Liberties Online*, 9 CDT POL'Y POST No. 3 (2003), http://www.cdt.org/publications/pp_9.03.shtml.

[18] Posting of Heather West to PolicyBeta, Regulations.gov Unleashes Wealth of Information for Users, http://blog.cdt.org/2008/01/15/regulationsgov-unleashes-wealth-of-information-for-users (Jan. 15, 2008); Regulations.gov, Welcome to the New Regulations.gov!, http://www.regulations.gov/fdmspublic/component/pubFooter_userTips (last visited Dec. 2, 2008).

[19] OpenRegulations.org, About This Site, http://www.openregulations.org/about/ (last visited Dec. 2, 2008).

[20] Farina et al., *supra* note 16, at 20.

[21] Farina et al., *supra* note 16, at 30

[22] Farina et al., *supra* note 16.

[23] H.R. REP. No. 109-153, at 138 (2006).

government users will not face higher costs if they encourage their stakeholders to make more extensive use of the system and streamlining the decision-making process for new features. If the ban on ancillary agency systems were also relaxed, the focus on structured, machine-readable data that we suggest here could be used to explore new functionality while still continuing to contribute documents to the existing Regulations.gov infrastructure.[24]

The tradeoff between standardization and experimentation, and the concerns about incomplete or inaccurate data in centralized government repositories such as Regulations.gov, are inherently difficult problems. USASpending.gov, created by legislation co-sponsored by Barack Obama and Tom Coburn in 2006,[25] presents another example: there, the desire to increase data quality by adopting a uniform method of identifying the recipients of federal funds has led to proposed amendments to the original legislation, aimed at improving data accuracy and standardization across agencies.[26] It is encouraging to see legislators take note of these intricate but significant details.

As long as government has a special role in the presentation and formatting of raw government data, certain desirable limits on what the government can do become undesirable limits on how the data can be presented or handled. The interagency group that sets guidelines for federal Web masters, for example, tells Web masters to manually check the status of every outbound link destination on their Web sites at least once each quarter.[27] And First Amendment considerations would vastly complicate, if not outright prevent, any effort to moderate online fora related to government documents. Considerations like these tend to make wikis, discussion boards, group annotation, and other important possibilities impracticable for government Web sites themselves.

Meanwhile, private actors have demonstrated a remarkably strong desire and ability to make government data more available and useful for citizens—often by going to great lengths to reassemble data that government bodies already possess but are not sharing in a machine-readable form. Govtrack.us integrates information about bill text, floor speeches and votes for both

---

[24] *See* Farina et al., *supra* note 16 (detailing specific steps toward a better Regulations.gov). These lie beyond the scope of our paper.

[25] Federal Funding Accountability and Transparency Act of 2006, Pub. L. No. 109-282, 120 Stat. 1186, *available at* http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_cong_public_laws&docid=f:publ282.109.pdf.

[26] Strengthening Transparency and Accountability in Federal Spending Act of 2008, S. 3077, 110th Cong., *available at* http://www.ombwatch.org/fedspending/ociiasintroduced.pdf.

[27] Web Managers Advisory Council, Establish a Linking Policy, http://www.usa.gov/webcontent/managing_content/organizing/links/policy.shtml (last visited Dec. 2, 2008).

houses of Congress by painstakingly reprocessing tens of thousands of Web pages.[28] It was created by a graduate student in linguistics in his spare time.[29] Carl Malamud, an independent activist, painstakingly took the SEC's data online[30] and is now attempting to open up judicial records,[31] which are currently housed behind subscription sites.

In some cases and to some degree, government bodies have responded to these efforts by increasing the transparency of their data. Key congressional leaders have expressed support for making their votes more easily available,[32] and the SEC is moving toward a format called XBRL that would increase the transparency of its own data.[33] In 2004, the OMB even asked that government units "to the extent practicable and necessary to achieve intended purposes, provide all data in an open, industry standard format permitting users to aggregate, disaggregate, or otherwise manipulate and analyze the data to meet their needs."[34] We argue below for a stronger impetus to provide open data: not "to the extent . . . necessary to achieve intended purposes", but as the main intended purpose of an agency's online publishing.

The federal government's current steps toward reusable data are valuable and admirable. But these efforts are still seen and prioritized as afterthoughts to the finished sites. As long as government bodies prioritize their own Web sites over infrastructures that will open up their data, the pace of change will be retarded.

II. INNOVATING FOR CIVIC ENGAGEMENT

Our goal is to reach a state where government provides all

---

[28] Govtrack.us: Tracking the U.S. Congress, http://www.govtrack.us (last visited Dec. 2, 2008).

[29] Govtrak.us, About Govtrack.us, http://www.govtrack.us/about.xpd (last visited Dec. 2, 2008).

[30] Posting of Taxpayer Assets, tap@essential.org, to listserver@essential.org, SEC's EDGAR on Net, What Happened and Why (Nov. 30, 1993, 10:36:34 EST), *available at* http://w2.eff.org/Activism/edgar_grant.announce.

[31] John Markoff, *A Quest to Get More Court Rulings Online, and Free*, N.Y. TIMES, Aug. 20, 2007, at C6.

[32] OMB Watch, Open House Project Calls for New Era of Access, http://www.ombwatch.org/article/articleview/3837/1/1 (last visited Feb. 2, 2009).

[33] *US SEC to Weigh XBRL Adoption Schedule on April 21*, REUTERS, Apr. 16, 2007, http://www.reuters.com/article/marketsNews/idUSN1642465120080416.

[34] Clay Johnson III, OFFICE OF MGMT. & BUDGET, EXECUTIVE OFFICE OF THE PRESIDENT, OMB MEMORANDUM: POLICIES FOR FEDERAL AGENCY PUBLIC WEBSITES 4 (2004), *available at* http://www.whitehouse.gov/omb/memoranda/fy2005/m05-04.pdf.

of its public data online[35] and there is vigorous third party activity to help citizens interact and add value to that data. Government need not—and should not—designate or choose particular parties to provide interaction. Instead, government should make data available to anyone who wants it, and allow innovative private developers to compete for their audiences.

A. Government Provides Data

Government should provide data in the form that best enables robust and diverse third party use. Data should be available, for free, over the Internet in open, structured, machine-readable formats to anyone who wants to use it. Using "structured formats" such as XML makes it easy for any third party service to gather and parse this data at minimal cost.[36] Internet delivery using standard protocols such as HTTP provides immediate real-time access to this data to developers. Each piece of government data, such as a document in XML format, should be uniquely addressable on the Internet in a known, permanent location.[37] This permanent address allows both third party services, as well as ordinary citizens, to link back to the primary unmodified data source as provided by the government.[38] All public data, in the highest detail available, should be provided in this format in a timely manner. As new resources are made available, government should provide data feeds, using open protocols such as RSS, to notify the public about the additions. These principles are consistent with the Open Government Working Group's list of

---

[35] Freedom of Information Act, 5 U.S.C. §552 (2002), *as amended by* Electronic Freedom of Information Act of 1996, Pub. L. No. 104-231, 110 Stat. 3048.

[36] To the extent that nontrivial decisions must be made about which formats to use, which XML schemas to use, and so on, government can convene public meetings or discussions to guide these decisions. In these discussions, government should defer to the reasonable consensus view of private site developers about which formats and practices will best enable development of innovative sites.

[37] Using the usual terms of art, the architectural design for data delivery must be RESTful. REST (short for Representational State Transfer) defines a set of principles that strives for increased scalability, generality, and data independence. The REST model adopts a stateless and layered client-server architecture with a uniform interface among resources. *See* Roy Thomas Fielding, Architectural Styles and the Design of Network-based Software Architectures (2000) (unpublished Ph.D. dissertation, University of California, Irvine), *available at* http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm.

[38] Concerns about data integrity—for example, possible modification by an intermediate service—can be addressed by using digital signatures. The originating Department or Agency can sign each primary source in such a way that data is verifiable and modification by an intermediary can be detected by the data recipient.

eight desirable properties for government data.[39]

In an environment with structured data, the politics of what to put on a home page are avoided, or made less important, because the home page itself matters less. And technical staff in government, whose hard work makes the provision of underlying data possible, will have the satisfaction of seeing their data used widely—rather than lamenting interfaces that can sometimes end up hiding valuable information from citizens.

B. Private Parties Present Data to Citizens

The biggest advantage of third party data processing is to encourage the emergence of more advanced features, beyond simple delivery of data. Examples of such features include

- *Advanced search:* The best search facilities go beyond simple text matching to support features such as multidimensional searches, searches based on complex and/or logical queries, and searches for ranges of dates or other values. They may account for synonyms or other equivalences among data items, or suggest ways to refine or improve the search query, as some of the leading Web search services already do.
- *RSS feeds:* RSS, which stands for "Really Simple Syndication," is a simple technology for notifying users of events and changes, such as the creation of a new item or an agency action. The best systems could adapt the government's own feeds (or other offerings) of raw data to offer more specialized RSS feeds for individual data items, for new items in a particular topic or department, for replies to a certain comment, and so on. Users can subscribe to any desired feeds, using RSS reader software, and those feeds will be delivered automatically to the user. The set of feeds that can be offered is limited only by users' taste for tailored notification services.
- *Links to information sources:* Government data, especially data about government actions and processes, often triggers news coverage and active discussion online. An information service can accompany government data with links to, or excerpts

[39] The group identified that government data must be complete, primary, timely, accessible, able to be processed by machines, non-discriminatory, non-proprietary and license-free. *See* Open Government Working Group, Open Government Data Principles, http://wiki.opengovdata.org/index.php/OpenDataPrinciples (last visited Dec. 2, 2008).

from, these outside sources to give readers context into the data and reactions to it.

- *Mashups with other data sources:* To put an agency's data in context, a site might combine that data with other agencies' data or with outside sources. For example, MAPlight.org combines the voting records of members of Congress with information about campaign donations to those members.[40] Similarly, the nonprofit group Pro Publica offers a map showing the locations of financial institutions that have received funds from the Treasury Department's Troubled Asset Relief Program (TARP).[41]
- *Discussion fora and wikis:* A site that provides data is a natural location for discussion and user-generated information about that data; this offers one-stop shopping for sophisticated users and helps novices put data in context. Such services often require a human moderator to erase off-topic and spam messages and to enforce civility. The First Amendment may make it difficult for government to perform this moderation function, but private sites face no such problem, and competition among sites can deter biased moderation.
- *Visualization:* Often, large data sets are best understood by using sophisticated visualization tools to find patterns in the data. Sites might offer users carefully selected images to convey these patterns, or they might let the user control the visualization tool to choose exactly which data to display and how.[42] Visualization is an active field of research and no one method is obviously best; presumably sites would experiment with different approaches.
- *Automated content and topic analysis:* Machine-learning algorithms can often analyze a body of data and infer rules for classifying and grouping data items.[43] By automating the classification of data, such models can aid search and foster analysis of trends.
- *Collaborative filtering and crowdsourced analysis:*

---

[40] Maplight.org., http://www.maplight.org (last visited Dec. 2, 2008).

[41] Pro Publica, Map: Show Me the TARP Money, http://www.propublica.org/special/bailout-map (last visited Jan. 12, 2009).

[42] "Many Eyes," for example, makes it simple for non-experts to dynamically visualize any custom dataset in a variety of different styles. Many Eyes, http://manyeyes.alphaworks.ibm.com/manyeyes/ (last visited Dec. 2, 2008).

[43] For example, software developed by Blei and Lafferty computed a topic model and classification of the contents of the journal *Science* since 1880. *See* David M. Blei & John D. Lafferty, *A Correlated Topic Model of Science*, 1 ANNALS OF APPLIED STAT. 17 (2007).

> Another approach to filtering and classification is to leverage users' activities. By asking each user to classify a small amount of data, or by inferring information from users' activities on the site (such as which items a user clicks), a site might be able to classify or organize a large data set without requiring much work from any one user.

Exactly which of these features to use in which case, and how to combine advanced features with data presentation, is an open question. Private parties might not get it right the first time, but we believe they will explore more approaches and will recover more rapidly than government will from the inevitable missteps. This collective learning process, along with the improvement it creates, is the key advantage of our approach. Nobody knows what is best, so we should let people try different offerings and see which ones win out.

For those desiring to build interactive sites, the barriers to entry are remarkably low once government data is conveniently available. Web hosting is cheap, software building blocks are often free and open source,[44] and new sites can iterate their designs rapidly. Successes thus far, including the Govtrack.us site that Joshua Tauberer built in his spare time,[45] show that significant resources are not required to enter this space. If our policy recommendations are followed, the cost of entry will be even lower.

III. PRACTICAL CONSIDERATIONS: HOW DO WE GET THERE FROM HERE?

Our proposal is simple: The new administration should specify that the federal government's primary objective as an online publisher is to provide data that is easy for others to reuse, rather than to help citizens use the data in one particular way or another.

The policy route to realizing this principle is to require that federal government Web sites retrieve their published data using the same infrastructure that they have made available to the public. Such a rule incentivizes government bodies to keep this infrastructure in good working order, and ensures that private parties will have no less an opportunity to use public data than the

[44] For example, the "LAMP stack," consisting of the Linux operating system, the Apache Web server, the MySQL database software, and the PHP scripting language, are available for free and widely used.

[45] About Govtrack.us, http://www.govtrack.us/about.xpd (last visited Dec. 2, 2008).

government itself does. The rule prevents the situation, sadly typical of government Web sites today, in which governmental interest in presenting data in a particular fashion distracts from, and thereby impedes, the provision of data to users for their own purposes.

Private actors have repeatedly demonstrated that they are willing and able to build useful new tools and services on top of government data, even if—as in the case of Joshua Tauberer's Govtrack.us[46] or Carl Malamud's SEC[47] and court document[48] initiatives—they have to do a great deal of work to reverse engineer and recover the structured information that government bodies possess, but have not published. In each case, the painstaking reverse engineering of government data allowed private parties to do valuable things with the data, which in turn created the political will for the government bodies (the SEC and Congress, in these cases) to move toward publishing more data in open formats.

When government provides reusable data, the practical costs of reuse, adaptation, and innovation by third parties are dramatically reduced. It is reasonable to expect that the low costs of entry will lead to a flourishing of third party sites extending and enhancing government data in a range of areas—rulemaking, procurement, and registered intellectual property, for example.

This approach could be implemented incrementally, as a pilot group of federal entities shift their online focus from finished Web sites to the infrastructure that allows new sites to be created. If the creation of infrastructure causes superior third party alternatives to emerge—as we believe it typically will—then the government entity can cut costs by limiting its own Web presence to functions such as branded marketing and messaging, while allowing third parties to handle core data interaction. If, on the other hand, third party alternatives to the government site do not satisfactorily emerge—as may happen in some cases—then the public site can be maintained at taxpayer expense. The overall picture is that the government's IT costs will decline in those areas where private actors have the greatest interest in helping to leverage the underlying data, while the government's IT costs will increase in those areas where, for whatever reason, there is no private actor in the world to step forward and create a compelling Web site based on the data. We expect that the former cases will easily outnumber the latter.

---

[46] Govtrack.us: Tracking the U.S. Congress, http://www.govtrack.us (last visited Dec. 2, 2008).

[47] U.S. Sec. & Exch. Comm'n., Electronic Data-Gathering, Analysis, and Retrieval (EDGAR) Database, http://www.sec.gov/edgar/searchedgar/webusers.htm (last visited Dec. 2, 2008).

[48] Markoff, *supra* note 31, at C6.

One key question for any effort in this area is the extent of flexibility in existing regimes. A number of recent laws have explicitly addressed the issue of putting government information on "Web sites." The E-Government Act of 2002, for example, asks each agency to put its contributions to the Federal Register, as well as various other information, on a public *Web site*.[49] This opens up a question of construal: Does an Internet location that contains machine-readable XML—which can be displayed directly in a Web browser and deciphered by humans but is designed to be used as input into a presentation system or engine—count as a "Web site"?[50]

If not, these statutory requirements may require government bodies to continue maintaining their own sites. It could be argued that XML pages are not Web pages because they cannot be conveniently understood without suitable software to "parse" them and create a human-facing display. But this objection actually applies equally and in the same way to traditional Web pages themselves: The plain text of each page contains not only the data destined for human consumption, but also information designed to direct the computer's handling or display of the underlying data, and it is via parsing and presentation by a browser program that users view such data.

One virtue of structured data, however, is that software to display it is easy to create. The federal government could easily create a general "government information browser" which would display any item of government information in a simple, plain, and universally accessible format. Eventually, and perhaps rapidly, standard Web browsers might provide such a feature, thereby making continued government provision of data browsing software unnecessary. Extremely simple Web sites that enable a structured data browser to display any and all government information may satisfy the letter of existing law, while the thriving marketplace of third party solutions realizes its spirit better than its drafters imagined.

We are focused in this paper on the government's role as a publisher of data, but it also bears mention that governmental bodies might well benefit from a similar approach to *collecting* data—user feedback, regulatory comments, and other official paperwork. This could involve private parties in the work of gathering citizen input, potentially broadening both the population

[49] E-Government Act of 2002, Pub. L. No. 107-347, 116 Stat. 2902, *available at* http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_public_laws&docid=f:publ347.107.pdf.

[50] Requirements that data be put "on the Internet" suffer no such ambiguities—providing the data in structured, machine-readable form on the Internet is sufficient to meet such a requirement.

from which input is gathered and the range of ways in which citizens are able to involve themselves in governmental processes. But it would raise a number of questions, such as the need to make sure that third party sites do not alter the data they gather before it reaches the government. These issues deserve further exploration but are beyond the scope of this paper.

IV. ALTERNATIVES AND COUNTERARGUMENTS

We argue that when providing data on the Internet, the federal government's core objective should be to build open infrastructures that enable citizens to make their own uses of the data. If, having achieved that objective, government takes the further step of developing finished sites that rely on the data, so much the better. Our proposal would reverse the current policy, which is to regard government Web sites themselves as the primary vehicle for the distribution of public data, and open infrastructures for sharing the data as a laudable but secondary objective.

The status quo has its virtues. As long as government Web sites themselves are the top priority, there is no risk that a lack of interest by private parties will limit citizens' access to government data. Instead, the government creates a system that every citizen can use (if not from home, then from a library or other public facility) without the need to understand the inner workings of technology. It might be argued that government ought to take a proprietary interest in getting its data all the way to individual citizens, and that relying on private parties for help would be a failure of responsibility. There is also a certain economy to the current situation: Under the current system, the costs of developing an open infrastructure for third party access are typically incurred in response to specific interest by citizens in accessing particular data—for example, Carl Malamud's campaign to move SEC data online.[51]

But, as described above, the status quo also has marked drawbacks. The institutional workings of government make it systematically incapable of adapting and improving Web sites as fast as technology itself progresses. No one site can meet as many different needs as well as a range of privately provided options can. And the idea that government's single site for accessing data will be a well-designed one is, as noted in Part I, optimistic at best. Moreover, the government already relies heavily on private parties for facilitating aspects of core civic activities—traveling to

[51] Posting of Taxpayer Assets, tap@essential.org, to listserver@essential.org, SEC's EDGAR on Net, What Happened and Why (Nov. 30, 1993, 10:36:34 EST), *available at* http://w2.eff.org/Activism/edgar_grant.announce.

Washington, calling one's representatives on the phone, or even going to the library to retrieve a paper public record all require the surrounding infrastructure within which the federal government itself is situated.

Another strategy—always popular in single-issue contexts—would be trying to "have our cake and eat it too" by fully funding *both* elaborate government Web sites and open data infrastructures. We have no quarrel with increasing the overall pool of resources available for federal Web development, but we do not think that any amount of resources would resolve the issue fully. At some point in each federal IT unit, there is apt to be someone who has combined responsibility for the full range of outward-facing Internet activities, whether these include an open infrastructure, a polished Web site, or both. Such people will inevitably focus their thoughts and direct their resources to particular projects. When open infrastructures drive Web sites, the infrastructure and site each rely on what the other is doing; it is extremely difficult to innovate on both levels at once.

Some people might want government to present data because they want access to the "genuine" data, unmediated by any private party. As long as there is vigorous competition between third party sites, however, we expect most citizens will be able to find a site provider they trust. We expect many political parties, activist groups, and large news organizations to offer, or endorse, sites that provide at least bare-bones presentation of government data. A citizen who trusts one of these providers or endorsers will usually be satisfied. To the extent that citizens want direct access to government data, they can access the raw data feeds directly. Private sites can offer this access, via the "permalinks" (permanent URLs) which our policy requires government-provided data items to have. If even this is not enough, we expect at least some government agencies to offer simple Web sites that offer straightforward presentation of data.

To the extent that government processes define standardized documents, these should be part of the raw data provided by the government, and should have a permanent URL. To give one example, U.S. patents should continue to be available, in standardized formats such as PDF, at permanent URLs. In addition, the Patent and Trademark Office should make the raw text of patents available in a machine-readable form that allows structured access to, for example, the text of individual patent claims.

Where it is necessary for a third party site to convince a user that a unit of government data is genuine, this can be

accomplished by using digital signatures.[52] A government data provider can provide a digital signature alongside each data item. A third party site that presents the data can offer a copy of the signature along with the data, allowing the user to verify the authenticity of the data item by verifying the digital signature without needing to visit the government site directly.

CONCLUSION

In this paper, we have proposed an approach to online government data that leverages both the American tradition of entrepreneurial self-reliance and the remarkable low-cost flexibility of contemporary digital technology. The idea, though it can be implemented in a comfortably incremental fashion, is ultimately transformative. It leads toward an ecosystem of grassroots, unplanned solutions to online civic needs.

Throughout the discussion, we have operated on the premise that citizen interaction with government data requires an intermediary: the federal government or, more effectively, third party innovators. In the long run, as the tools for interacting with data continue to improve and become increasingly intuitive, we may reach a state in which citizens themselves interact directly with data without needing any intermediary.

The federal government's current Web presence falls far short of what is possible. The energy and opportunity for change that comes with President Obama's electoral victory could help end the current system of episodic upgrading of government Web sites whereby sites will continue to drift out of date. If the administration instead steps forward to adopt the grassroots model we suggest, then the federal government's Internet presence will be *permanently* improved—citizen access to government data will keep pace with technology's progress indefinitely into the future.

[52] Digital signatures are cryptographic structures created by one party (the "signer") that can be verified by any other party (the "verifier") such that the verifier is assured that the signature could only have been created by the signer (or someone who stole the signer's secret key), and that the document to which the signature applies has not been altered since it was signed. *See, e.g.*, NAT'L INST. OF STANDARDS & TECH., U.S. DEP'T OF COMMERCE, FIPS PUB No.186-2, DIGITAL SIGNATURE STANDARD (DSS) (2000), *available at* http://csrc.nist.gov/publications/fips/fips186-2/fips186-2-change1.pdf.