

Government Data and the Invisible Hand

David Robinson*, Harlan Yu*[†], William Zeller*[†] and Edward W. Felten*^{†‡}
Contact: {dgr, harlanyu, wzeller, felten}@princeton.edu

1 Introduction

If the next Presidential 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 websites 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 websites, as the core of its online publishing responsibility.

In the current Presidential cycle, all three candidates have indicated that they think the federal government could make better use of the Internet. Barack Obama's platform explicitly endorses "making government data available online in universally accessible formats."¹ Hillary Clinton, meanwhile, remarked that she wants to see much more government information online.² John McCain, although expressing excitement about the Internet, has allowed that he would like to delegate the issue, possibly to a vice-president.³

But the situation to which these candidates are 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. The federal government has shown itself consistently unable to keep pace with the fast-evolving power 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 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 websites themselves use the same open systems for accessing the underlying data as they make available to the public at large**.

*Center for Information Technology Policy, Princeton University

[†]Department of Computer Science, Princeton University

[‡]Woodrow Wilson School of Public and International Affairs, Princeton University

¹Obama for America, *Technology and Innovation for a New Generation*, available at <http://www.barackobama.com/issues/technology> (accessed April 16, 2008).

²Hillary Clinton, 'Meet the Press' Transcript for Jan. 13, 2008, available at <http://www.msnbc.msn.com/id/22634967> (accessed April 16, 2008).

³John McCain, *Part II: CNN/YouTube Republican Presidential Debate Transcript*, available at <http://www.cnn.com/2007/POLITICS/11/28/debate.transcript.part2/index.html> (accessed April 18, 2008).

Our approach follows the engineering principle of separating data from interaction, which is commonly used in constructing websites.⁴ Government must provide data, but we argue that websites 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 offer 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 with its vibrant marketplace of engineering ideas to discover what works.

2 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 badly out of date that seven senators, including John McCain, 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 a web site whose basic structure has remained unchanged since 2001.⁷ Experts report that in order to make any practical use of the system, 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

Regulations.gov, a government-wide docket publishing system launched in 2003 and used by “nearly all Departments and Agencies,” faces similar trouble.¹¹ Originally designed to open up the federal rulemaking process and broaden the population submitting comments, it was launched with a limited search engine 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.¹² It did not become plausibly able to fulfill its original mission

⁴Most sophisticated websites 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 websites already use such a separation internally. Government sites that currently separate these functions are already partway to the goal we espouse.

⁵Library of Congress, *About THOMAS*, available at http://thomas.loc.gov/home/abt_thom.html (accessed April 16, 2008).

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

⁷Internet Archive Wayback Machine for <http://www.fcc.gov> from September 17, 2001, available at <http://web.archive.org/web/20010917033924/http://www.fcc.gov/> (accessed April 18, 2008).

⁸See Cynthia Brumfield, *The FCC is the Worst Communicator in Washington* (2007), available at http://www.ipdemocracy.com/archives/002640the_fcc_is_the_worst_communicator_in_washington.php, and Jerry Brito, *FCC.gov: The Docket that Doesn't Exist*, available at <http://techliberation.com/2007/11/01/fccgov-the-docket-that-doesnt-exist/> (accessed April 16, 2008).

⁹Jerry Brito, *Hack, Mash & Peer: Crowdsourcing Government Transparency* (2007), available at <http://ssrn.com/abstract=1023485>, at 5.

¹⁰Jerry Brito, *FCC.gov: Searching in Vain* (2007), available at <http://techliberation.com/2007/10/29/fccgov-searching-in-vain> (accessed April 16, 2008).

¹¹Regulations.gov, *What is on this Site*, available at http://www.regulations.gov/search/this_site.jsp (accessed April 16, 2008).

¹²Center for Democracy and Technology, *CDT Policy Post* (2003) Vol. 9, No. 3 available at <http://www.cdt.org/publications/>

**Do NOT cite. A final version of this essay will appear in Volume 11
of the Yale Journal of Law and Technology in Fall 2008.**

until five years later, when the site was relaunched with a better, more intuitive search engine.¹³

Once it had invested money and time in developing a new and better version of its own site, Regulations.gov also released its underlying data in a computer-readable format (in this case, RSS) which allowed any interested person or group to create an alternative, enhanced version of the website.¹⁴ This has led to the creation of OpenRegulations.org, which competes with Regulations.gov by offering “pared down, simple-to-navigate listings of new agency dockets.”¹⁵

The reasons for Regulations.gov’s slow development are not mysterious. It was designed by committee, subject to repeated delays, and constrained by a funding cutoff after a Congressional panel found 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.”¹⁶

These problems are unfortunately typical of government web efforts. An online compliance checklist for designers of government websites identifies no fewer than 24 different regulatory regimes with which all public government web sites must comply.¹⁷ Ranging from privacy and usability to FOIA compliance to the requirements of the Paperwork Reduction Act and, separately, the Government Paperwork Elimination Act, each of these requirements alone is reasonable enough. They reflect the considered judgment of our political process, informed by whatever understanding of information technology was available when they were written. But the stultifying cumulative effect of these rules has not been, and probably would not be, endorsed by anyone.¹⁸ 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.¹⁹ Moreover, 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 webmasters, for example, tells webmasters to manually check the status of every outbound link destination on their websites at least once each quarter.²⁰ And First Amendment considerations would vastly complicate, if not outright prevent, any effort to moderate online fora related to government documents. Considerations like these make wikis, discussion boards, group annotation, and other important possibilities impracticable for government websites 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 decline to share. Govtrack.us integrates information about bill text,

pp_9.03.shtml (accessed April 16, 2008).

¹³Center for Democracy and Technology, *Regulations.gov Unleashes Wealth of Information for Users* (2008), available at <http://blog.cdt.org/2008/01/15/regulationsgov-unleashes-wealth-of-information-for-users> (accessed April 16, 2008).

¹⁴Regulations.gov, *Welcome to the New Regulations.gov!*, available at http://www.regulations.gov/fdmspublic/component/pubFooter_userTips (accessed April 16, 2008).

¹⁵OpenRegulations.org, *About this site*, available at <http://www.openregulations.org/about/> (accessed April 16, 2008).

¹⁶Cindy Skrzycki, *Document Portal Sticks on Funding*, *The Washington Post*, Jan. 10, 2006, at D01.

¹⁷Web Content Managers Advisory Council, *Requirements Checklist for Government Web Managers*, available at http://www.usa.gov/webcontent/reqs_bestpractices/reqs_checklist.shtml/ (accessed April 16, 2008).

¹⁸For example, many requirements mandate certain content to be included on homepages. 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.

¹⁹And 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., *The Knight Open Government Survey* (2007), available at http://www.knightfoundation.org/research_publications/detail.dot?id=221378, at 7.

²⁰Web Managers Advisory Council, *Establish a Linking Policy*, available at http://www.usa.gov/webcontent/managing_content/organizing/links/policy.shtml (accessed April 18, 2008).

floor speeches and votes for both houses of Congress by painstakingly reprocessing tens of thousands of webpages.²¹ It was created by a graduate student in linguistics in his spare time.²² Carl Malamud²³, an independent activist, painstakingly took the SEC's data online²⁴ and is now attempting to open up judicial records²⁵, 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,²⁶ and the SEC is moving toward a format called XBRL that would increase the transparency of its own data.²⁷ In 2004, the Office of Management and Budget 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."²⁸ 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.

These limited steps, and the general goal, are admirable. But they are still seen and prioritized as afterthoughts to the finished sites. As long as government bodies prioritize their own websites over infrastructures that will open up their data, the pace of change will be retarded.

3 Innovating For Civic Engagement

Our goal is to reach a state where government provides all public data²⁹ and there is vigorous third party activity to help citizens interact and add value to it. 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.

3.1 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.³⁰ 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

²¹Govtrack.us: Tracking the U.S. Congress, *available at* <http://www.govtrack.us> (accessed April 18, 2008).

²²About Govtrack, *available at* <http://www.govtrack.us/about.xpd> (accessed April 18, 2008).

²³media.org, *Encapsulation of Carl Malamud*, *available at* <http://media.org/carl.html> (accessed April 18, 2008)

²⁴Taxpayer Assets Project, *SEC's EDGAR on Net, What Happened, and Why*, *available at* http://w2.eff.org/Activism/edgar_grant. announce (accessed April 18, 2008)

²⁵John Markoff, *A Quest to Get More Court Rulings Online, and Free*, *The New York Times*, Aug. 20, 2007

²⁶OMB Watch, *Open House Project Calls for New Era of Access*, May 15, 2007, *available at* <http://www.ombwatch.org/article/articleview/3837/1/1?TopicID=1> (accessed April 18, 2008)

²⁷Reuters, *US SEC to weigh XBRL adoption schedule on April 21*, Apr. 16, 2007.

²⁸Clay Johnston III, *OMB Memorandum: Policies for Federal Agency Public Websites* (2004), *available at* <http://www.whitehouse.gov/omb/memoranda/fy2005/m05-04.pdf>.

²⁹5 U.S.C. §552. (The Freedom of Information Act of 1966, amended by the Electronic Freedom of Information Act of 1996.)

³⁰To 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.

in XML format, should be uniquely addressable on the Internet in a known, permanent location.³¹ 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.³² 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 eight desirable properties for government data.³³

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. 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.

3.2 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 best web search services do.
- *RSS feeds*: RSS 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 will offer 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 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

³¹Using 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 (Doctoral Dissertation)* (2000), available at <http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>.

³²Concerns 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.

³³The group identified that government data must be complete, primary, timely, accessible, machine-processable, non-discriminatory, non-proprietary and license-free. See Open Government Working Group, *Open Government Data Principles*, available at <http://wiki.opengovdata.org/index.php/OpenDataPrinciples> (accessed April 16, 2008).

**Do NOT cite. A final version of this essay will appear in Volume 11
of the Yale Journal of Law and Technology in Fall 2008.**

records of members of Congress with information about campaign donations to those members.³⁴ Similarly, the Sunlight Foundation offers a map showing the locations of Congressional earmarks.³⁵

- *discussion forums 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. 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.³⁶ By automating the classification of data, such models can aid search and foster analysis of trends.
- *collaborative filtering and crowdsourced analysis*: 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 specifically 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 from the inevitable missteps. This collective learning process—and the improvements it creates—is the key advantage of our approach. Nobody knows what is best, so we should let people try different approaches and see which one wins 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 source³⁷ and new sites can iterate their designs rapidly. Successes thus far, such as Josh Tauberer's building Govtrack in his spare time,³⁸ 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.

³⁴Maplight.org. Money and Politics: Illuminating the Connection, available at <http://www.maplight.org> (accessed April 18, 2008).

³⁵Sunlight Foundation, *Earmark Map*, available at <http://www.sunlightlabs.com/earmarks> (accessed April 18, 2008).

³⁶For example, software developed by Blei and Lafferty computed a topic model and classification of the contents of the journal *Science* since 1890. See David M. Blei and John D. Lafferty, *A Correlated Topic Model of Science* (2007), *Annals of Applied Statistics*, 1:1 at 17-35.

³⁷For 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 are widely used.

³⁸About Govtrack, available at <http://www.govtrack.us/about.xpd> (accessed April 18, 2008).

4 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 websites retrieve the underlying 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 government itself does. The rule prevents the situation, sadly typical of government websites 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 Govtrack³⁹ or Carl Malamud's SEC⁴⁰ and court document⁴¹ initiatives—they have to do a great deal of work to reverse engineer and recover the structured information that government bodies have, 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 websites 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 website based on the data. We expect that the former cases will easily outweigh the latter.

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 “websites.” 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 *website*.⁴² 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

³⁹Govtrack.us: Tracking the U.S. Congress, *available at* <http://www.govtrack.us> (accessed April 18, 2008).

⁴⁰Electronic Data-Gathering, Analysis, and Retrieval (EDGAR) Database for the U.S. Securities and Exchange Commission, *available at* <http://www.sec.gov/edgar/searchedgar/webusers.htm>.

⁴¹John Markoff, *A Quest to Get More Court Rulings Online, and Free*, The New York Times, Aug. 20, 2007

⁴²44 U.S.C. §101

as a “website”?⁴³

If not, these statutory requirements may require government bodies to continue maintaining their own sites. It could be argued that XML pages are not webpages because they cannot be 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 webpages 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 websites 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—through an open infrastructure. This would be, in computer science parlance, an “API” (application programming interface). This involves private parties in the work of gathering citizen input, potentially broadening both the population 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.

5 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 websites 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 websites 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

⁴³Requirements 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.

example, Carl Malamud's campaign to move SEC data online.⁴⁴

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 Section 2, optimistic at best. Moreover, the government already relies heavily on private parties for facilitating aspects of core civic activities—traveling to 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 websites 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 website, or both. Such people will inevitably focus their thoughts and direct their resources to particular projects. When open infrastructures drive websites, 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, 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 websites 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, e.g., the text of individual patent claims.

Where it is necessary for a citizen to convince a third party that a unit of government data is genuine, this can be accomplished by using digital signatures.⁴⁵ 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.

⁴⁴Taxpayer Assets Project, *SEC's EDGAR on Net, What Happened, and Why*, available at http://w2.eff.org/Activism/edgar_grant.announce (accessed April 18, 2008)

⁴⁵Digital signatures are cryptographic structures created by one party (the “signer”) that can be verified by any other party (the “verifier”) such that that 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., National Institute of Standards and Technology, *Federal Information Processing Standard 186: Digital Signature Standard*, (2000) available at <http://csrc.nist.gov/publications/fips/fips186-2/fips186-2-change1.pdf> (accessed April 18, 2008).

6 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 a new President could easily lead to an episodic upgrading of government web sites, a sudden shift after which sites will continue to drift out of date. If the administration instead steps forward and adopts 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.