Testimony of Jim Harper
Director of Information Policy Studies
The Cato Institute
to the
Committee on Oversight & Government Reform
United States House of Representatives
at a hearing entitled
“Addressing Transparency in the Federal Bureaucracy: Moving Toward A More Open Government”
March 13, 2013

Executive Summary

President Obama’s 2008 campaign helped light a fire under the government transparency movement that still burns. However, the effort to produce transparent government has flagged. This is essentially because of poor awareness of exactly what practices produce transparent government.

Confusion between “open government” and “open government data” illustrates this. They are often treated as interchangeable, but the first is about revealing the deliberations, management, and results of government, and the second is general availability of data that the government has produced, covering any subject matter.

More importantly, the transparency community has failed to articulate what it wants. A quartet of data practices would foster government transparency: authoritative sourcing, availability, machine-discoverability, and machine-readability. The quality of government data publication by these measures is low.

We are not waiting for the government to produce good data. At the Cato Institute, we have begun producing data ourselves, starting with legislation that we are marking up with enhanced, more revealing XML code.

Our efforts are hampered by the unavailability of fundamental building blocks of transparency, such as unique identifiers for all the organizational units of the federal government. There is today no machine-readable organization chart for the federal government.

Well-published data, such as what the DATA Act requires, would allow the transparency community to propagate information about the government in widely varying forms to a public that very much wants to understand what happens in Washington, D.C.

Chairman Issa, Ranking Member Cummings, and members of the committee:

Thank you for the opportunity to testify before you today. I am keenly interested in the subject matter of your hearing, and I hope that my testimony will shed some light on your oversight of federal government transparency and assist you in your deliberations on how to promote this widely agreed-upon goal.

My name is Jim Harper, and I am director of information policy studies at the Cato Institute. Cato is a non-profit research foundation dedicated to preserving the traditional American principles of limited government, individual liberty, free markets, and peace. In my role there, I study the unique problems in adapting law and policy to the information age, issues such as privacy, intellectual property, telecommunications, cybersecurity, counterterrorism, and government transparency.

For more than four years, I have been researching, writing on, and promoting government transparency at Cato. For more than a dozen years, I have labored to provide transparency directly through a Web site I run called WashingtonWatch.com. Other transparency-related work of mine includes serving on the Board of Directors of the National Priorities Project, serving on the Board of Advisors of the Data Transparency Coalition, and serving on the Advisory Committee on Transparency, a project of the Sunlight Foundation run by my co-panelist today Daniel Schuman.

WashingtonWatch.com is still quite rudimentary and poorly trafficked compared to sites like Govtrack.us, OpenCongress, and many others, but collectively the community of private, non-profit and for-profit sites have more traffic and almost certainly provide more information to the public about the legislative process than the THOMAS Web site operated by the Library of Congress and other government sites.

There is nothing discreditable about THOMAS, of course, and we appreciate and eagerly anticipate the improvements forthcoming on Congress.gov. But the many actors and interests in the American public will be best served by looking at the federal government through many lenses—more and different lenses than any of us can anticipate or predict. Thus, I recommend that you focus your transparency efforts not on Web sites or other projects that interpret government data for the public. Rather, your task should be to make data about the government’s deliberations, management, and results available in the structures and formats that facilitate experimentation. There are dozens—maybe hundreds—of ways the public might examine the federal government’s manifold activities.

Delivering good data to the public is no simple task, but the barriers are institutional and not technical. Your leadership, if well-focused, can produce genuine progress.

I will try to illustrate how to think about transparency by sharing a short recent history of transparency, a few reasons why the transparency effort has flagged, the publication practices that will foster transparency, our work at the Cato Institute to show the way, the need for a machine-readable government organization chart, and finally the salutary results that the DATA Act could have for transparency.

A Short Recent History of Federal Government Transparency

President Obama deserves credit for lighting a fire under the government transparency movement in his first campaign and in the first half of his first term. To roars of approval in 2008, he sought the presidency making various promises that cluster around more open, accessible government. Within minutes of his taking office on January 20, 2009, the Whitehouse.gov website declared: "President Obama has committed to making his administration the most open and transparent in history."[1] And his first presidential memorandum, entitled "Transparency and Open Government," touted transparency, public participation, and collaboration as hallmarks of his forthcoming presidential administration.[2]

In retrospect, the prediction of unparalleled transparency was incautiously optimistic. But at the time, the Obama campaign and the administration's early actions sent strong signals that energized many communities interested in greater government transparency.

My own case illustrates. In December 2009, between the time of President Obama's election and his inauguration, I hosted a policy forum at Cato entitled: "Just Give Us the Data! Prospects for Putting Government Information to Revolutionary New Uses."[3] Along with beginning to explore how transparency could be implemented, the choice of panelists at the event was meant to signal that agreement on transparency would cross ideologies and parties, regardless of differences over substantive policies. That agreement has held.

In May 2009, White House officials announced on the new Open Government Initiative blog that they would elicit the public's input into the formulation of its transparency policies.[4] The public was invited to join in with the brainstorming, discussion, and drafting of the government's policies.

The conspicuously transparent, participatory, and collaborative process contributed to an "Open Government Directive," issued in December 2009 by Office of Management and

[1] Macon Phillips, "Change Has Come to Whitehouse.gov," The White House Blog, January 20, 2009 (12:01 p.m. EDT), http://www.whitehouse.gov/blog/change_has_come_to_whitehouse-gov.
[2] Barack Obama, "Transparency and Open Government," Presidential Memorandum (January 21, 2012), http://www.whitehouse.gov/the-press-office/transparency-and-open-government.
[3] Cato Institute, "Just Give Us the Data! Prospects for Putting Government Information to Revolutionary New Uses," Policy Forum, December 10, 2008, http://www.cato.org/event.php?eventid=5475.
[4] Jesse Lee, "Transparency and Open Government," May 21, 2009, http://www.whitehouse.gov/blog/2009/05/21/transparency-and-open-government.

Budget head Peter Orszag.[5] Its clear focus was to give the public access to data. The directive ordered agencies to publish within 45 days at least three previously unavailable "high-value data sets" online in an open format and to register them with the federal government's data portal, data.gov. Each agency was to create an "Open Government Webpage" as a gateway to agency activities related to the Open Government Directive.

They did so with greater or lesser alacrity.

But while pan-ideological agreement about transparency has held up well, the effort to produce transparent government has flagged. The data.gov effort did not produce great strides in government transparency or public engagement. And many of President Obama's transparency promises went by the wayside.

His guarantee that health care legislation would be negotiated "around a big table" and televised on C-SPAN was quite nearly the opposite of what occurred.[6] His promise to post all bills sent him by Congress online for five days was nearly ignored in the first year.[7] His promise to put tax breaks online in an easily searchable format was not fulfilled. Various other programs and projects have not produced the hoped-for transparency, public participation, and collaboration. And the Special Counsel to the President for Ethics and Government Reform, who handled the White House's transparency portfolio, decamped for an ambassadorial post in Eastern Europe at the mid-point of President Obama's first term.

It's easy (and cheap) for critics of the president to chalk his transparency failures up to campaign disingenuousness or political calculation. It is true that the Obama administration has not shone as brightly on transparency as the president promised it would. But my belief is that transparency did not materialize in President Obama's first term because nobody knew what exactly produces transparent government. The transparency community had not put forward clearly enough what it wanted from the government, and the transparency effort got sidetracked in a subtle but important way from "open government" to "open government data."

Open Government vs. Open Government Data

When the White House instructed agencies to produce data for data.gov, it gave them a very broad instruction: produce three "high-value data sets" per agency. According to the open government memorandum:

[5] Peter R. Orszag, "Memorandum for the Heads of Executive Departments and Agencies, Subject: Open Government Directive," M 10-06, December 8, 2009, http://whitehouse.gov/open/documents/open-government-directive [hereinafter "Open Government Directive"].

[6] "Negotiate Health Care Reform in Public Sessions Televised on C-SPAN," Politifact.com, http://www.politifact.com/truth-o-meter/promises/obameter/promise/517/health-care-reform-public-sessions-C-SPAN/.

[7] Jim Harper, "Sunlight Before Signing in Obama's First Term," Cato blog, February 12, 2013, http://www.cato.org/blog/sunlight-signing-obamas-first-term.

> High-value information is information that can be used to increase agency accountability and responsiveness; improve public knowledge of the agency and its operations; further the core mission of the agency; create economic opportunity; or respond to need and demand as identified through public consultation.[8]

That's a very broad definition. Without more restraint than that, public choice economics predicts that the agencies will choose the data feeds with the greatest likelihood of increasing their discretionary budgets or the least likelihood of shrinking them. That's data that "further[s] the core mission of the agency" and not data that "increase[s] agency accountability and responsiveness."

"It's the Ag Department's calorie counts," as I wrote before the release of data.gov data sets, "not the Ag Department's check register."[9] And indeed that's what the agencies produced.

In a grading of the data sets, I found that most failed to expose the deliberations, management, and results of the agencies. Instead, they provided data about the things they did or oversaw. The Agriculture Department produced data feeds about the race, ethnicity, and gender of farm operators; feed grains, "foreign coarse grains," hay, and related commodities; and the nutrients in over 7,500 food items.

"That's plenty to chew on," I wrote in my review of all agency data sets, "but none of it fits our definition of high-value."[10]

The agencies, and the transparency project, were diverting from open government to open government data. David Robinson and Harlan Yu identified this shift in policy focus in their paper: "The New Ambiguity of 'Open Government.'" They wrote:

> Recent public policies have stretched the label "open government" to reach any public sector use of [open] technologies. Thus, "open government data" might refer to data that makes the government as a whole more open (that is, more transparent), but might equally well refer to politically neutral public sector disclosures that are easy to reuse, but that may have nothing to do with public accountability.[11]

[8] Open Government Directive.

[9] Jim Harper, "Is Government Transparency Headed for a Detour?" Cato blog, January 15, 2010, http://www.cato.org/blog/government-transparency-headed-detour.

[10] Jim Harper, "Grading Agencies' High-Value Data Sets," Cato blog, February 5, 2010, http://www.cato.org/blog/grading-agencies-high-value-data-sets.

[11] Harlan Yu and David G. Robinson, "The New Ambiguity of 'Open Government,'" UCLA Law Review 59, no. 6 (August 2012): 178.

There's nothing wrong with open government data, but the heart of the government transparency effort is getting information about the functioning of government. I think in terms of a subject-matter trio that I have mentioned once or twice already—deliberations, management, and results.

Data about these things are what will make for a more open, more transparent government. That is what President Obama campaigned on in 2008, it is what I believe you are interested in producing through your efforts, and it is what I believe will satisfy the American public's demand for transparency. Everything else, while entirely welcome, is just open government data.

Publication Practices for Transparent Government

Deliberations, management, and results are complex processes, so it is important to be aware of another, more technical level on which the transparency project got bogged down. The transparency community did not meet public demand for, and political offer of, government transparency with a clear articulation of what produces it. We failed to communicate our desire for well-published, well-organized data, making clear also what that is.

Believing this to be the problem, I embarked in 2010 on a mission to learn what data publication practices will produce government transparency. A surprisingly intense, at times philosophical, series of discussions with propeller-heads of various types—information scientists, librarians, data geeks, and so on—allowed me to meld their way of seeing the world with what I knew of public policy processes.

In the Cato report, "Publication Practices for Transparent Government" (attached to my testimony as Appendix I), I sought to capture four categories of data practice that can produce transparency: authoritative sourcing, availability, machine-discoverability, and machine-readability. I summarized them briefly as follows:

> The first, authoritative sourcing, means producing data as near to its origination as possible—and promptly—so that the public uniformly comes to rely on the best sources of data. The second, availability, is another set of practices that ensure consistency and confidence in data.
>
> The third transparent data practice, machine-discoverability, occurs when information is arranged so that a computer can discover the data and follow linkages among it. Machine-discoverability is produced when data is presented consistent with a host of customs about how data is identified and referenced, the naming of documents and files, the protocols for communicating data, and the organization of data within files.

> The fourth transparent data practice, machine-readability, is the heart of transparency, because it allows the many meanings of data to be discovered. Machine-readable data is logically structured so that computers can automatically generate the myriad stories that the data has to tell and put it to the hundreds of uses the public would make of it in government oversight.[12]

Following these data practices does not produce instant transparency. Users of data throughout the society would have to learn to rely on governmental data sources. Transparency, I wrote,

> turns on the capacity of the society to interact with the data and make use of it. American society will take some time to make use of more transparent data once better practices are in place. There are already thriving communities of researchers, journalists, and software developers using unofficial repositories of government data. If they can do good work with incomplete and imperfect data, they will do even better work with rich, complete data issued promptly by authoritative sources.[13]

Our efforts have not ceased with describing how the government can publish data to foster transparency. Starting in January 2011, the Cato Institute began working with a wide variety of groups and advisers to "model" governmental processes as data and then to prescribe how this data should be published.

Our November 2012 report, "Grading the Government's Data Publication Practices"[14] (part of which is attached to my testimony as Appendix II) examined how well the government publishes data reflecting legislative process and the budgeting, appropriating, and spending processes. Having broken down each element of these processes, we polled the community of government data users to determine how well that data is produced, and we issued letter grades.

The grades were generally poor, and my assessment (mine alone, not endorsed by other participants in our process) was that the House has taken a slight lead on government transparency, showing good progress with the small part of government it directly controls. The Obama administration, having made extravagant promises, lags the House by comparison. Since the release of the report, more signs of progress have come from the House, including forthcoming publication of committee votes, for example.[15] This

[12] Jim Harper, "Publication Practices for Transparent Government," Cato Institute Briefing Paper no. 121, September 23, 2011, http://www.cato.org/publications/briefing-paper/publication-practices-transparent-government [hereinafter "Publication Practices"].

[13] *Id.*

[14] Jim Harper, "Grading the Government's Data Publication Practices," Cato Policy Analysis no. 711, November 5, 2012, http://www.cato.org/publications/policy-analysis/grading-governments-data-publication-practices.

[15] *See* Jim Harper, "Sunlight Before Signing in Obama's First Term," Cato blog, February 12, 2013, http://www.cato.org/blog/sunlight-signing-obamas-first-term.

gap could easily be closed, however, if the administration gives focused attention to data transparency.

We Are Extending and Enriching Government Data Publication

Having assessed the publication practices that we believe will foster transparency, and having graded the government's publication practices in key areas, we are not waiting for good data to materialize. We have begun producing the data ourselves.

The low-hanging fruit for government transparency is the legislative process. In Congress, the long existence of the THOMAS Web site and the practice of publishing bills in a data format called XML (eXtensible Markup Language) make it easier to track what is happening than it is in other areas. But it is not easy enough, and we are working to make it even easier.

At the Cato Institute, we have acquired and modified software that allows us to extend the XML markup in existing bills. While most of the code embedded in the bills that Congress produces deals with the appearance of the bills when printed, we are adding code that fleshes out what the bills mean.

Using the data modeling we have done, we are tagging references to existing law in an organized, machine-readable way, so that people can learn instantly when a provision of law they care about is the subject of a bill. We are tagging budget authorities—both authorizations of appropriations and appropriations themselves—so that proposals to expend taxpayer funds are instantly and automatically available to the public and to you in Congress.

This being Sunshine Week, we are holding sessions tomorrow and Friday to examine how our enhanced bill XML can be a tool for Wikipedians. People across the country go to Wikipedia for information, including information about public affairs, and we would like to see that they are met there by good information about prominent pieces of legislation and our laws.

We plan to take our experience with marking up bills to other types of government documents and other processes. But it is difficult work. In the bills you write, you in Congress refer to existing law in varied, sometimes anachronistic ways. The varying ways your bills denote budget authorities sometimes make it very hard to represent clearly how many dollars are being made available for how many years.

But one of the problems we really should not be having is identifying the organizational units of government referred to in bills. In addition to tagging existing law and proposed spending, we tag agencies, bureaus, and such. But we are essentially unable to tag entities below the agency and bureau level, and the tagging we are doing uses identifiers we cannot be sure are reliable.

We are doing what we can to make bills available for computer interpretation—and we should be able to do wonders when appropriation season comes around—but we are hindered by the lack of a machine-readable federal government organization chart. We need this basic government data, which is essential to transparency.

Needed: A Machine-Readable Federal Government Organization Chart

Data is a collection of abstract representations of things in the world. We use the number "3," for example, to reduce a quantity of things to an abstract, useful form—an item of data. Because clerks can use numbers to list the quantities of fruits and vegetables on hand using numbers like "3," for example, store managers can effectively carry out their purchasing, pricing, and selling instead of spending all of their time checking for themselves how much of everything there is. Data makes everything in life a little easier and more efficient for everyone

Legislative and budgetary processes are not a grocery store's produce department, of course. They are complex activities involving many actors, organizations, and steps. The Cato Institute's modeling of these processes reduced everything to "entities," each having various "properties." The entities and their properties describe the things in legislative and budgetary processes and the logical relationships among them, like members of Congress, the bills they introduce, hearings on the bills, amendments, votes, and so on.

A member of Congress is an important entity in legislative process, as you might imagine. And happily, there are already systems in place to identify them accurately to computers. The "Biographical Directory of the United States Congress" is a compendium of information about all present and former members of the U.S. Congress (as well as the Continental Congress), including delegates and resident commissioners. The "Bioguide" website at bioguide.congress.gov is a great resource for searching out historical information about members.

Bioguide does a brilliant thing in particular for making the actions of Members of Congress machine-readable. It assigns a unique ID to each of the people in its database.

To illustrate how Bioguide works, I've copied the Bioguide IDs for each member of this committee into a table. The Bioguide IDs you see in this table are used across machine-readable documents and government Web sites to make crystal clear to computers exactly whom is being referred to when the name of a member of Congress is used, no matter what variation there is in the way the member is referred to in the resource.

This simple idea, of providing unique IDs for important components of governmental processes, is a basic building block of government transparency. Having Bioguide IDs has vastly improved the public's ability to oversee Congress, and the Congress's ability to track its own actions.

House Committee on Oversight and Government Reform: Membership and Bioguide IDs

Republican Members	Bioguide ID	Democratic Members	Bioguide ID
Rep. Darrell E. Issa	I000056	Rep. Elijah Cummings	C000984
Rep. John L. Mica	M000689	Rep. Carolyn Maloney	M000087
Rep. Michael Turner	T000463	Rep. Eleanor Holmes Norton	N000147
Rep. John J. Duncan	D000533	Rep. John Tierney	T000266
Rep. Patrick T. McHenry	M001156	Rep. Wm. Lacy Clay	C001049
Rep. Jim Jordan	J000289	Rep. Stephen Lynch	L000562
Rep. Jason Chaffetz	C001076	Rep. Jim Cooper	C000754
Rep. Tim Walberg	W000798	Rep. Gerald Connolly	C001078
Rep. James Lankford	L000575	Rep. Jackie Speier	S001175
Rep. Justin Amash	A000367	Rep. Matt Cartwright	C001090
Rep. Paul Gosar	G000565	Rep. Mark Pocan	P000607
Rep. Pat Meehan	M001181	Rep. Tammy Duckworth	D000622
Rep. Scott DesJarlais	D000616	Rep. Danny K. Davis	D000096
Rep. Trey Gowdy	G000566	Rep. Peter Welch	W000800
Rep. Blake Farenthold	F000460	Rep. Tony Cardenas	C001097
Rep. Doc Hastings	H000329	Rep. Steve Horsford	H001066
Rep. Cynthia Lummis	L000571	Rep. Michelle Lujan Grisham	L000580
Rep. Rob Woodall	W000810		
Rep. Thomas Massie	M001184		
Rep. Doug Collins	C001093		
Rep. Mark Meadows	M001187		
Rep. Kerry Bentivolio	B001280		
Rep. Ron DeSantis	D000621		

But unique identification has not been applied to many other parts of government. The most glaring example is the lack of authoritative and unique IDs for the organizational units of government. The agencies, bureaus, programs, and projects that make up the executive branch of government are not uniquely identified to the public in a similar way, and the relationships among all the federal government's organizational units is not authoritatively published anywhere.

In short, there is no machine-readable federal government organization chart. This is a glaring problem and a serious impediment to government transparency.

Were there a machine-readable federal government organization chart, the unique identifiers for organizational units could appear in all manner of document: budgets, authorization bills, appropriation bills, regulations, budget requests, and so on. Then we could use computing to help knit together stories about all the different agencies in our federal government, what they do, and how they use national resources. Internal management and congressional oversight would both strengthen. The DATA Act holds out the possibility of all this happening.

The DATA Act: That Organization Chart and More

The DATA Act essentially requires there to be a machine-readable government organization chart and much more. Building on widely lauded experience of the Recovery Accountability and Transparency Board, the DATA Act calls for data reporting standards that are "widely accepted, non-proprietary, searchable, platform-independent [and] computer-readable."[16] This is the centerpiece of the DATA Act, from my perspective, and it is true of versions of the bill in both the House and the Senate last Congress.

To be a success, such standards must not only uniquely identify all the organizational units that carry out Congress's instructions in the executive branch. They must also identify budget documents; legislation; budget authorities; warrants, apportionments, and allocations; obligations; non-federal parties; and outlays.

Having unique identifiers for each of these things, and attributes that signal their relationships to one another, will allow vast stores of information to emerge from the data. "Seeing" the relationship between a given budget, a given appropriations bill, the obligation it funded, and an outlay of funds will make available the "story" of what Congress does year in and year out with taxpayers' money.

This data will make internal and congressional oversight far stronger. And it may help knit together the entire budget and spending process, so that expenditures can be matched to the results that Congress sought when it created programs and funded them. You in Congress and your constituents in the public will have better awareness of what happens in Washington, D.C. and in government offices around the country.

All this serves goals that span partisan and ideological lines. Organizing the spending process will reduce waste, fraud, and abuse in the first instance, as the likelihood of discovery will rise. Debates about programs may base themselves less on ideology and more on actual statistics about what spending achieved what results. In my "Publication Practices" paper, I wrote:

> Transparency is likely to produce a virtuous cycle in which public oversight of government is easier, in which the public has better access to factual information,

[16] H.R. 2146 (112 Cong., 2nd Sess.).

> in which people have less need to rely on ideology, and in which artifice and spin have less effectiveness. The use of good data in some areas will draw demands for more good data in other areas, and many elements of governance and public debate will improve.[17]

I do believe this is true, though these ideal outcomes will not be reached automatically. Indeed, they will require a lot of effort to achieve.

Essential to producing the standards that foster these benefits is the existence of one authority positioned to require them. It seems natural for spending data standardization to be handled by the Office of Management and Budget, but that office has so far proven unwilling to move forward. Thus, the creation of a Federal Accountability and Spending Transparency board or commission may be warranted. My preference, of course, would be for economy in the creation of more federal entities to track…

I was surprised in September of 2011 to see the Congressional Budget Office estimate for the version of the DATA Act this committee reported to the House. The estimate of $575 million in outlays to implement the DATA Act over five years was quite nearly unbelievable. The thing that may make it believable is if waste, fraud, and abuse infects implementation of the DATA Act.

I believe that it will cost less than the CBO predicts to implement the DATA Act should it become law. Modifying federal data systems may have costs in the short term, but complying with standards should have essentially no cost after the initial retooling. If it does take as much to fully implement the Act as the CBO estimates, that is proof of a sort that we need oversight systems like this that can hold costs down.

The GRANT Act, FOIA Reform, and More

Our transparency research and work has not extended to federal grant-making, which is a significant subset of all federal spending. It seems obvious that bringing transparency and organizational rigor to grant administration would have similar salutary effects to what we can expect in government spending generally.

Outright waste would be curtailed. The results of grant-making for public policy goals would be clearer. And participants in the grant-making process would be more sure of fair treatment.

I understand there are concerns with the version of the GRANT Act introduced in the last Congress, such as with the potential that anonymous peer review might be undercut by transparency. This is a genuine issue, which can almost certainly be overcome with some careful thinking and planning. If it cannot, my belief is that the interest of the taxpayer in

[17] Publication Practices.

accountable grant administration is generally superior to the interests of peer reviewers in privacy or anonymity.

Of course, Freedom of Information Act reforms are an important part of the transparency agenda. I am not expert in FOIA, and it is my hope that proactive and thorough data transparency may partially diminish the need for FOIA requests because transparency policy has made clear what deliberative processes agencies have used, for example.

Even in a world with the fullest data transparency, there will be a public need for access to key government documents and information on request. I support the FOIA reforms that will get the most important information disseminated the most broadly so that American democracy functions better and so that public oversight of the government is strong.

Conclusion

It is a pleasure to work on an issue like transparency, widely supported as it is across partisan and ideological lines. Transparency is a means to various ends that can co-exist. I believe, for example, along with my conservative friends, that transparency will reduce the demand for government and increase the demand for private authority over decisions and spending that the government currently controls. If transparency produces this result, it will be a product of democratic processes that I think my liberal and progressive friends would be hard-pressed to reject. If, on the other hand, transparency wrings waste, fraud, and abuse out of government programs, validating them and increasing their support, I will enjoy the gain of having a better-managed government.

If there is division in the transparency issue, it is between the outsiders and the insiders. Information is power, and non-transparent practices are a way of preserving power for the few who have attained it.

The enjoyment of power by the few is inconsistent with the underlying theory of democracy, of course, and with our shared American commitment to the idea that power springs from the people. The moral high-ground in debates about transparency is always with those who want to know more about what their government is doing with their money and their rights. While there may be some narrow exceptions to the rule that the people have a right to know, the transparency status quo is far from that line.

Anything this committee can do to improve the quality and quantity of data about the government’s deliberations, management, and results will bring credit to the committee and this Congress.

Appendix I

Publication Practices for Transparent Government

CATO INSTITUTE BRIEFING PAPERS

Publication Practices for Transparent Government

by Jim Harper

No. 121 September 23, 2011

Executive Summary

Government transparency is a widely agreed upon goal, but progress on achieving it has been very limited. Transparency promises from political leaders such as President Barack Obama and House Speaker John Boehner have not produced a burst of information that informs stronger public oversight of government. One reason for this is the absence of specifically prescribed data practices that will foster transparency.

Four key data practices that support government transparency are: authoritative sourcing, availability, machine-discoverability, and machine-readability. The first, authoritative sourcing, means producing data as near to its origination as possible—and promptly—so that the public uniformly comes to rely on the best sources of data. The second, availability, is another set of practices that ensure consistency and confidence in data.

The third transparent data practice, machine-discoverability, occurs when information is arranged so that a computer can discover the data and follow linkages among it. Machine-discoverability is produced when data is presented consistent with a host of customs about how data is identified and referenced, the naming of documents and files, the protocols for communicating data, and the organization of data within files.

The fourth transparent data practice, machine-readability, is the heart of transparency, because it allows the many meanings of data to be discovered. Machine-readable data is logically structured so that computers can automatically generate the myriad stories that the data has to tell and put it to the hundreds of uses the public would make of it in government oversight.

Jim Harper is director of information policy studies at the Cato Institute and webmaster of government transparency website WashingtonWatch.com.

Digitization and the Internet have had transformative effects on bookselling, banking and payments, news, and entertainment, but these technologies have barely touched government.

Introduction

> I'll make our government open and transparent, so that anyone can ensure that our business is the people's business.
>
> When there's a tax bill being debated in Congress, you will know the names of the corporations that would benefit and how much money they would get.
>
> The Internet offers new opportunities to open the halls of Congress to Americans in every corner of our nation.
>
> The lack of transparency in Congress has been a problem for generations, under majorities Republican and Democrat alike. But with the advent of the Internet, it's time for this to change.

During electoral and political campaigns, transparency promises seem to flow like water. The quotes above—the first two from President Obama and the second two from Speaker Boehner—were issued during these officials' runs for higher office. Then-senator Barack Obama (D-IL) spoke about transparency to roars of applause on the presidential campaign trail.[1] Minority Leader John Boehner (R-OH), seeking to outflank Speaker Nancy Pelosi and the Democrats on their management of the House of Representatives, touted transparency in a video recorded in the U.S. Capitol's Statuary Hall.[2]

So what happens to transparency promises when the campaign ends? Having achieved their political goals, do elected officials just throw transparency out like so much bathwater? Digitization and the Internet have had transformative effects on bookselling, banking and payments, news, and entertainment, but these technologies have barely touched government. This might be consistent with the predictions of public choice economics: transparency will generally reduce politicians' freedom of action by increasing public oversight. Having more information available to more people would allow more second-guessing of politicians' decisions, weakening inputs into electoral success such as fundraising and logrolling. So maybe politicians will always reject transparency, even as they sing its praises.

But the story is more complex than that. If transparency promises were convenient election-eve fibs, Obama would probably not have made issuing an open government memorandum his first executive action upon taking office. With his election only months past and a re-election campaign nearly as far away as it could be, he called for a transparent, participatory, and collaborative federal government on his first day in office.[3] Late in Obama's first year, his director of the Office of Management and Budget (OMB), Peter Orszag, issued an Open Government Directive instructing executive departments and agencies to take specific actions to implement the principles of transparency, participation, and collaboration.[4] The White House created an "Open Government Initiative" page on its website, Whitehouse.gov,[5] and documented the work on its open-government blog.[6] Pursuant to the Orszag directive, agencies produced "open government plans" and released "high-value data sets," registering the latter on the new Data.gov website.[7] These actions do not reflect insincerity, but rather a good-faith effort to advance transparency goals.

Boehner commands far fewer organs of government than the president, but his efforts, and those of the Republican House leadership, have been roughly proportional to the president's. Upon taking control in the 112th Congress, Republicans passed a package of rule changes aimed at increasing transparency.[8] This package included a 72-hour rule requiring the posting of bills "in electronic form" for three days before a vote on the House floor. In April, Boehner and Majority Leader Eric Cantor (R-VA) wrote a letter to the House Clerk asking her to transition toward publishing legislative data in open formats.[9]

Like Obama, House Republicans are following up their transparency promises with efforts that are at least adequate. All probably recognize that transparency is a growing demand of the public and that meeting that demand will help them win elections. Yet neither the administration nor Congress has become notably more transparent.

Perhaps the transparency shortage can be explained by simple lack of effort. Time constraints exist for politicians just like everyone else—if they spent more time on transparency, we would probably get more of it. But this conclusion is too facile and not revealing enough. It provides no way forward other than to join the interest-group scrum urging "more dedication" to a particular cause. And it offers no hope of resolving the problem: How will we know when we've got transparency?

The better explanation for transparency floundering in the face of good-faith effort is indeterminacy. Though transparency is a widely recognized value, nobody knows exactly what it is. The steps that produce transparent government are opaque—ironically—so transparency efforts have not crystallized or produced positive change.

Though transparency is a widely recognized value, nobody knows exactly what it is.

The Data.gov project helps to illustrate this. The OMB's Open Government Directive called for each agency to publish three high-value data sets. According to the memorandum, high-value information is:

> . . . information that can be used to increase agency accountability and responsiveness; improve public knowledge of the agency and its operations; further the core mission of the agency; create economic opportunity; or respond to need and demand as identified through public consultation.[10]

For all its verbiage, that definition has almost no constraints. Anything could be ranked "high-value." And sure enough, agencies' high-value data feeds ran the gamut from information that might truly inform the public to things that could interest only the tiniest niche researcher. An informal Cato Institute analysis examined the data streams each agency released and graded the agencies using a more-demanding definition of high value: whether their releases provide insight into agency management, deliberations, or results.[11] There were some As, but Ds were more common. The rating given to the Agriculture Department is an example of the latter:

> The Ag Department produced data feeds about the race, ethnicity, and gender of farm operators; feed grains, "foreign coarse grains," hay, and related items; and the nutrients in over 7,500 food items. That's plenty to chew on, but none of it fits our definition of high-value.

"Management, deliberation, and results" is only a loose description of what information the public might most benefit from seeing, and agencies were not obligated by OMB to rise to that standard, so a poor grade is not damning. More discussion between the public (represented by the transparency community) and government will specify more concretely what information should be published.

But there are more questions than this: How is it that thousands of data feeds are supposed to "connect up" with the websites, researchers, and reporters who would turn them into useful information? How is it that a great mass of data is supposed to find the people that can use it, and the people find the data?

In December 2008, a Cato Institute policy forum focused on the transparency commitments of the new president. Its title was "Just Give Us the Data! Prospects for Putting Government Information to Revolutionary New Uses."[12] The Obama administration did exactly that, publishing lots and lots of data, but transparency did not flourish. The simple sloganeer's demand for "the data" was immature.

In this paper, we explore more deeply how to produce government transparency. Transparency is not only about access to data, or its substance in management, deliberation, or results. Government transparency is a set of data-publication practices that facilitate "finding"—the matching up of information with public interest.

Recognizing the discrete publication practices that produce transparency can crystallize the forward progress that everyone wants in this area. Rather than "more effort," or other indeterminate demands, the transparency community and the public can measure whether government entities and agencies are publishing data consistent with transparency. Measurable transparency behaviors will help the public hold officials to account after their transparency promises have brought them into office. Government officials should know that the public is not satisfied, and will not be satisfied, until data flows like water and government information like a mighty stream.

Publication Practices for Transparent Government

Water is a useful metaphor for data. Salt water can't quench a person's thirst. Nor can a block of ice, or water vapor. Water has to be in a specific form, liquid and reasonably pure, for it to be drinkable. So it is with government data and transparency. There is an endless sea of publications, websites, speeches, news reports, data feeds, and social media efforts, but somehow the public still thirsts for information it can use. Water, water, everywhere, and not a drop to drink.

Information must be delivered in specific ways—"liquid" and relatively "pure"—for the body politic to consume it well.

It turns out that information, like water, must be delivered in specific ways—"liquid" and relatively "pure"—for the body politic to consume it well. Data about government agencies, entities, and activities must be published in particular ways if it is going to facilitate transparency.

When the Republican 104th Congress created the THOMAS legislative system in 1995, it was a huge advance for transparency—a huge advance from a very low baseline, at least. Publication on THOMAS might be summarized as a disclosure model, in which certain key documents and records were made available "as is," or in a limited number of forms optimized for the World Wide Web, which is just one way of sharing information on the Internet. Much of the discussion today about putting bills online and having members of Congress "read the bill" is still framed in terms of disclosure, but the underlying demand is something more.

Since the mid-90s, the way people use the Internet has changed dramatically. "Web 2.0" is the buzzword that captures the shift from one-way publishing toward interactivity and user-generated content. On the modern Internet, data serves as a platform for interaction and decisionmaking.

The next steps in government transparency must match this change, going beyond simple disclosure of documents and records to publication of data in ways the modern Internet can use. Governments should publish data that reflects their deliberations, management, and results in highly accessible ways that natively reveal meaning. Publication of government data this way will allow the public to digest government information and take concrete actions in response.

Four categories of information practice, discussed below, are a foundation for government transparency that the public is quickly coming to expect. They are: authoritative sourcing, availability, machine-discoverability, and machine-readability.

A number of papers and documents produced over the last few years have advocated, described, and discussed transparent government data practices in parallel to these concepts. A 2007 working group meeting in Sebastopol, California, for example, produced a suite of 8 principles for open government data,[13] which was later increased to 10 principles in August, 2011.[14] The recommendations of the Open House Project, also published in 2007, were animated by these good information practices.[15] There are many other such documents.[16]

The federal government has not embraced these data publication practices yet, so transparency has not yet flourished as it could. In part, this is because the specific information practices that will set the stage for transparency are still unclear.

Everyone knows what drinkable water is, but it takes physicists, chemists, and biologists to make sure drinkable water is what comes out of the tap. Parallel sciences go into producing data in formats that are consistent, fully useful, and fully informative. The discussion that follows does not fully detail each information practice that will foster government transparency, but it should alert people familiar with computing and the Internet to the practices that prepare data adequately for public consumption.

The digital world is different from the physical world in many ways. Data can come and go in ways that physical things do not, so things that are given, obvious, or easy in the physical world have to be thought through and watched after in the digital world. For this reason, the first transparent data practice—establishment of "authority" around data—requires unique attention.

Authoritative Sourcing

Just as people look to authoritative books or thinkers to know the right answers about science, life, or philosophy, they look to authority in data to be confident of having the right information and a fully accurate account of the things data describe. Authority in data is a lot like authority in other areas—it is about knowing where to look for data and what sources to trust. Because of people's willingness to trust and use reliable resources more than unreliable ones, data can be more or less transparent depending on the quality of its authority.

Authority means a number of related concepts dealing with who is responsible for publication and who is recognized as responsible. The word "authoritative" has a couple of senses, both of which are relevant to authoritative sourcing. One sense is formal: data should come from the authoritative source—which is almost always the entity that creates or first captures the data.[17] Uniting the data and its origin is a good idea because authoritative sourcing reduces the chance of error and fraud, for example. Authoritative sourcing also makes it easier for newcomers to find data, because the creator and the publisher are the same. The shortest possible "chain of custody" between the information's origination and its publication is best.

If the data's creator delegates the responsibility to publish, then the second sense of authoritative is in play. That is the sense that some entity is recognized by the relevant public as fully reliable. The delegated publisher should be recognized as the authoritative data source.

It is sometimes easiest to illustrate good practices by highlighting error. A small gap in authority exists today in the publication of certain U.S. federal legislative data, such as the text of bills. Congress has delegated the authority to publish information about bills and their texts to the Government Printing Office, which puts such information on its FDsys website.[18] But if you were to ask most experienced Washington hands, and even many people working with legislative data, what the source of legislative information was, they would probably think first of the Library of Congress' THOMAS system.[19] But THOMAS is a downstream republisher of data, some of which the Government Printing Office originates on behalf of the Congress. Most users of legislative data do not look to FDsys or THOMAS, however. They use data collections at govtrack.us,[20] a website whose operator curates legislative data for public use.

These small gaps in authority are not a significant problem. But multiple sources publishing the same data without revealing its provenance can be a problem for authority. The entity that has the legal authority

Data can be more or less transparent depending on the quality of its authority.

to publish data and the entity that is recognized by the relevant public as the authoritative source should be the same.

A practice that promotes authority is real-time or near-real-time publication.[21] If an agency like the Department of Defense, for example, were to publish a compilation of contract documents every month, rather than a real-time, hourly, or daily record of such documents, then data aggregators, lobbying firms, news outlets, or others might make a good business of collecting contract information and publishing it before the Defense Department does. Various audiences, hungry for information, would rightly turn to these organizations and divide their loyalties among data sources. Though meeting a legitimate need, this dynamic would produce multiple nonauthoritative data sources, introducing inefficiency and the potential for error and confusion—as well as literal delay—into the process. These are all things that weaken transparency.

Authoritative sourcing—the notion of one entity known to have responsibility for publishing data—is a simple but important transparency practice.

The authority required for transparency is *earned* through prompt publication of data in useful open standards—“authority through being awesome,” in the words of the Sunlight Foundation’s Eric Mill.[22] This contrasts with the assertion of authority that exists when the focus is on publishing in file formats that explicitly include authority information. Digital mechanisms that seek to ensure authenticity, such as cryptographically signed files, certainly have their place in securing against forgery, for example. But ensuring authenticity this way can be counterproductive to transparency if it slows publication or locks data in difficult-to-use formats.

Transparency will also be strengthened if an authority has ways to correct data.[23] Especially in widely variable human processes like legislating and regulating, there are plenty of opportunities for incorrect data to see publication. This highlights the need for an authoritative publisher. When the authority becomes aware of error—and it should be open to receiving such information from data users—the authority can publish the fix and propagate the newly corrected information to all downstream users.

If several data sources act as originators for downstream users, errors may persist in some systems while they are corrected in others. The information produced by one set of data may be different from another, sowing confusion and detracting from transparency’s goals. Society would waste time and effort in the absence of good authority determining which data set is right, rather than moving forward on the things that make life better for people.

Authoritative sourcing—the notion of one entity known to have responsibility for publishing data—is a simple but important transparency practice. It is an anchor for the next set of transparency-friendly data publication practices, clustered around availability.

Availability

Availability consists of a variety of practices that ensure information can reliably be found and used.[24] Availability in the digital world is a lot like availability in the physical world—it’s having access to what you need—but availability is very easy to violate in the data realm. A physical thing, like a phone booth, takes a fair amount of work to make unavailable, so we don’t think about the importance of availability with such things. Data can be made unavailable with careless planning or the touch of a button, so availability is important to plan for. Availability has a number of features.

Permanence is an important part of availability.[25] A thing is not truly available unless it exists for good. Data that reflects the activities of an agency in issuing regulations, for example, reflects very important real-world activity. Just as society needs a permanent record of this lawmaking process to have confidence in it, data users need a permanent record of data to be confident in the data they use and the results it produces. Once published, data should exist forever, so that one person can confirm another's version of events, so that anyone can check the original data source, and so on. Data that disappears at some point after publication is harder to rely on. Part of making data available is keeping it available forever.

Similarly, data should be stable, meaning it should always be found in the same location. Think of whether you might consider a pay phone to be available for your use if it was only sometimes on the street corner near your office. If a pay phone moved from place to place at random times, it would be hard to know if you could actually use it at any given hour. It would not be fully available. It is the same with data, which has to be in the same place all the time to be truly available.

Data is available when it is complete.[26] A partial record is partial because some part of it is unavailable. That is not sufficient, because users of the data could produce incorrect results with incomplete information. Of course, any data set must have a scope. But if the scope is not obvious from context, it should be explained in the data's documentation. A partial record is unreliable, and it cannot be used to tell the stories that full data records can, so it does not foster transparency as it should.

In general, data about government deliberation, management, and results should be made available on the Internet for free.[27] If government entities are executing well on authoritative publication, this practice should have no costs additional to the creation of the data. Execution of key government functions, creation of data about that execution, and publication of that data should all be essentially the same thing. Data that is not at the core of governmental functions or other exceptions—gigantic, niche-interest, or rarely used data sets, for example—might be made available on other terms. But cost-free online access to essential-government-function data is best.

The processes by which data is made available are also relevant. Data is fully available when it is available both in bulk and incrementally. In bulk means that the entire data set is available all at once. This is so that a new user can access the data or existing users can double-check that a copy of the data they have is accurate and complete. Incremental means that updates to the data are published in a way that allows a user to update his or her copy of the data. Requiring users to download bulk data just to access recent changes may be prohibitively costly, so it does not fully meet the need for data availability.

There is another sense to availability—a legal sense. In fact, there are two senses to legal availability. Data is fully available when it is structured using standards that are unencumbered by intellectual property claims.[28] There are techniques for manipulating and storing data that are covered by patent claims, for example. To use them, one must pay the owner of the patent a licensing fee. If it costs money to use the standard in which data is published, that data is not fully available. It is encumbered by licensing costs.

Similarly, data itself may sometimes be subject to intellectual property claims. If a string of text in a database is copyrighted, for example, that datum is not fully available. It is encumbered by legal claims that limit its use. This will not usually be the case with federal government data; works of the government are not generally copyrightable. But some materials that are made a part of government records may be copyrightable or copyrighted, and some government entities may claim copyright in their documents or try to assert other forms of restriction on information they produce or publish.[29] Government data should not be controlled by intellectual property laws or otherwise restricted, and data that is so controlled is not sufficiently available.

"Available" in the world of data is more complex than it sounds. There are a variety of ways that data can be rendered unavailable, so it is important to think about availability and to provide it in support of transparency. With authoritative sources making

Availability consists of a variety of practices that ensure information can reliably be found and used.

data available, machine-discoverability and machine-readability round out the data publication practices that can produce transparency.

There must be sufficient order to the way things are referred to in links and data for that data to be truly machine-discoverable.

Machine-Discoverability

As we move more deeply into the technical details of transparency, we come to a concept closely related to availability, but going more to the particular techniques by which data is made available. This is machine-discoverability. The question here is whether data is arranged so that a computer can discover the data and follow linkages among it.

In a literal sense, data is machine-discoverable when it can be found by a machine. Because of powerful consensus around protocols, this basically means using hypertext transfer protocol (HTTP), the language used behind all websites,[30] and links using hypertext markup language (HTML)[31] that direct machines to data.

But full machine discoverability means more than following these two customs alone: it means following a host of customs about how data is identified and referenced, including the organization and naming of links, the naming of files, the protocols for communicating files, and the organization of data within files. There must be sufficient order to the way things are referred to in links and data for that data to be truly machine-discoverable.

A consistent uniform resource locator (URL) structure is an important way of making data discoverable. The links from the home page of a website to substantive data should exist and make sense. The words in the link, and the links themselves, should be accurately descriptive or orderly in some other logical way to help people find things. Just as people follow links they think will take them to the data they want, search engines "spider" data—crawling, spiderlike, through every link they find—to record what data is available.

One illustration of discoverability failure comes from early implementation of Obama's "Sunlight Before Signing" promise on Whitehouse.gov. As a campaigner, Obama promised he would post bills online for five days prior to signing them. When the White House began to implement this practice early in the new administration, it began putting pages up on Whitehouse.gov for bills Congress had sent to the president. But these pages were not within the link structure that starts on the Whitehouse.gov homepage. A person (or search engine) following every link on Whitehouse.gov would not have arrived at these pages.[32] The bills were literally posted on the Whitehouse.gov domain, but they were not discoverable in any practical sense. The only way to find them was to use Whitehouse.gov's search engine, knowing ahead of time what terms to search for.

Sometimes machine-discoverability will be thwarted by the failure to publish like data in like ways. In 2007, Congress began requiring its members to disclose the earmarks that they had requested from the appropriations committees. This was an important step forward for transparency—some disclosure is better than none—but nothing about the disclosure rules made the information machine-discoverable. Members of Congress put their disclosures on their own websites with no consistency as to how the files were named. The result was that earmark requests were still hard to find—for humans and machines both. Members of Congress followed the path of least resistance, which also happened to frustrate transparency and the small transfer of power to the public that transparent publication would have produced. Fully transparent earmark disclosure would have required earmark requests to be consistently linked or, more likely, to have been reported to a central clearinghouse for publication, such as the appropriations committees receiving the requests.

Not only was the dispersion of earmark data across websites a problem, it was also in multiple, inconsistent file formats. Some members posted their information on webpages in HTML format. Some posted portable-document file (PDF) lists of their earmarks. Still others posted scanned PDF images of earmark request printouts. Because there was no consistency among the earmark disclosures, computers had a very hard time recognizing them as being similar, and earmark transparency was weakened. To enhance public access to earmark information, transparency and taxpayer groups gathered earmark data from all over the House and Senate websites.[33] Though these assemblages lacked authority, they were more transparent than the undiscoverable earmark request webpages produced pursuant to House and Senate rules.

File naming, storage, and transfer conventions are important. When they look at a file, some machines (and a few people) look at the name of the file to figure out how to open it and learn what it contains. There are strong conventions about file naming that help machines do this—conventions that are familiar to many. Webpages often end with .html, for example. Microsoft Word files end with the suffix .doc. Excel files end with .xls. Simple text files, or plain text, end with .txt. HTTP improves on file-name extensions by indicating files' multipurpose Internet mail extension (MIME) type, which is independent of file name extensions.[34]

When these customs are violated it makes data harder to discover by machine. The Federal Election Commission (FEC), for example, has created its own class of text file that it labels .fec.[35] This means that a visitor does not know what kind of files they are. The FEC site serves files using file transfer protocol (FTP), which does not signal the MIME type. This frustrates a computer scan or search-engine spider's attempt to open the files. Worst of all, the files are zipped, meaning they have been compressed using an algorithm that makes it hard for a Web crawler to look inside them.

Ultimately, discoverability is a function of how easy or hard it is for machines to locate data. Various good practices make data more discoverable, and failure to follow these practices makes it less discoverable. These things have to be thought through in the data world, which does not have the same fixity that makes maps reliable in the physical world.

Machine-discoverability is the product of relatively mechanical practices and conventions about data publication—"where things are on the Internet." But as it reaches higher levels of refinement, discoverability of files and their content blends in with what might be called *conceptual* discoverability—"what the things on the Internet are." Data is most discoverable is if its meaning is apparent from its structure and organization. This blends into machine-readability, which allows data, once discovered, to see substantive use.

Machine-Readability

Machine-readability is what truly brings data to life and makes it transparent.

Machine-readability is what truly brings data to life and makes it transparent. Machine-readability goes beyond the generic finding in machine-discoverability to a deeper level—a level at which the data can be used in meaningful and valuable ways.[36] As legislative data guru Josh Tauberer writes, "[D]ata's value depends not only on its subject, but also on the format in which the information is shared. Format determines the value of the resource and the extent to which the public can exploit it for analysis and reuse."[37] The Association for Computing Machinery puts it similarly: "Data published by the government should be in formats and approaches that promote analysis and reuse of that data."[38] Analysis and reuse—that means searching, sorting, linking, and transforming information in ways that support people's substantive goals.

Machine-readable data has what might be called semantic richness. That means that *meaning* is easy to discover from it. Transparency is meant to give the public access to the meaning of various government actions the way the public has access to meaning in other areas of life.

The human brain brings a wealth of semantic information to bear when it perceives the world. When a student sitting in an American history class, for example, hears another student talk about Wilson, she knows from the context of the situation that the other student is probably talking about the former president of the United States. A student in a popular-film class might assume Wilson to be the name of the volleyball friend of Tom Hanks in the movie *Castaway*. A student in a physical education course might assume Wilson to be the company that makes volleyballs and tennis balls. To say these people know these things is to say that they make quick—blindingly quick—calculations about what the word "Wilson" refers to when they hear it.

A computer does not do those kinds of calculations unless it is told to do them. To make computers comprehend strings of letters like "Wilson," these strings have to be disambiguated, or normalized. That is, they have to be placed into a logical structure, often using distinct identifiers that substitute for clumsy identifiers like names. This allows machines to recognize distinctions among things that are otherwise similar.

Distinct Identifiers

Like Wilson, the name Rogers has many meanings. It's the name of a telecommunications company in Canada. It's also a city in Arkansas, and another city in Minnesota. It's a county in Oklahoma, and it's the name of a famous architect. A man and his wife in Portland, Oregon, are named Rogers—as are their three children—and lots of other people around the country. While the name Rogers does a lot of good in small circles to distinguish among people, it is a terrible way in to find a specific person or thing in the big digital world. Even the custom of attaching a given name to a surname doesn't work in digital environments. Just ask Mike Rogers.

Mike Rogers is the name of two different people currently serving in the House of Representatives. One Mike Rogers is from Michigan and the other Mike Rogers is from Alabama. Their staffs undoubtedly receive mail and phone calls meant for the other Mike Rogers all the time. But Congress has done something important to clear up this ambiguity. It has disambiguated these Mike Rogerses (and all elected representatives) within its Bioguide system.[39]

Mike Rogers, the representative of Michigan's 8th district, has the Bioguide ID: "R000572." Mike Rogers, the representative of Alabama's 3rd district, has the Bioguide ID: "R000575." Substituting abstract strings of letters and numbers for names helps computers identify more accurately the information they are scanning. With a Bioguide lookup table, a computer can tell when data refers to Mike Rogers from Michigan and when it refers to Mike Rogers from Alabama. It will never mistake these Rogerses for any other Mike Rogers, much less the famous architect or the Canadian telecommunications company.

This is how the structuring of data gives it semantic meaning. With broadly known and well-followed naming conventions like this, information about Mike Rogers and every other member of Congress can easily and quickly be collected and shared with their constituents and the public as a whole.

This type of structure can be applied to all generic entities in a data system, allowing computers to observe the logical relationships among them and to tell relevant stories automatically. When data properly disambiguates representatives' names, their votes, and party affiliation, for example, computers can easily calculate party cohesion from one vote to another. If vote data includes the date, as it should, computers can quickly calculate party cohesion over time. If representatives' names and Bioguide IDs are correlated to states (as they are), computers can automatically calculate state and regional cohesion in voting. Each addition of data expands the range of stories the data can tell.

There are just a few small illustrations of the literally thousands of different stories that computers might generate automatically from disambiguated or normalized data. There are dozens of different entities involved in legislative processes, dozens

There are literally thousands of different stories that computers might generate automatically from disambiguated or normalized data.

more in budgeting and appropriations, dozens more in regulatory processes, litigation, and so on. There are many overlaps among the entities involved in each of these, and relationships among them as well. For transparency to flourish, all these entities must be described in data with logical coherence.[40]

Formatted Data

When data is published in machine-readable ways, its meanings can come to life, and it can be the foundation of truly transparent government. The ways this can be done have many layers of complexity, but they are worth understanding in general. Most people are familiar with formats, the agreed-upon arrangements, protocols, and languages used to collect, store, and transmit data. From the moment information is captured digitally—when a word is typed on a computer keyboard or a camera and microphone record a speech—it is arranged and rearranged through various formats that convert it to binary data (ones and zeroes, or on/off, up/down). This binary data can later be converted back into letters and words, symbols, and the combinations of sounds and images that comprise audio and video.

Just as there are formats for collecting, storing, and transmitting data, there are formats for organizing data in ways that optimize it for human consumption. Some of the most familiar and easiest to understand are in the area of typesetting and display.

If an author means to emphasize a certain point, and makes a word or phrase display as boldface text to do that, her word processing software will record that display preference. (“Only **fourteen** people in Peoria drive a Fiat Spider!”) Later copies of the document should retain signals that make her chosen words appear in bold. When the text is converted to the format suitable for the World Wide Web—hypertext markup language, or HTML—the signal that the word “fourteen” should be displayed bold looks like this:

```
Only  <b>fourteen</b>  people
in Peoria drive a Fiat Spider!
```

When a browser like Internet Explorer or Firefox sees the signals <b> and </b>, it displays the material between the “start” and “end” signals as bold. A human looking at the resulting text knows that the author wanted to convey the importance of the word “fourteen.”

This is a very rudimentary example, and it deals only with display and printing. The same technique could be used for highlighting semantic information in a machine-readable way. For example, the words “Fiat Spider” could be surrounded by signals that indicate a discussion about automobiles:

```
Only <b>fourteen</b> people in
Peoria drive a <car make=″Fiat″
model=″Spider″>Fiat   Spider
</car>!
```

This uses the same kind of signaling to allow a properly programmed computer to recognize that this is a discussion of cars, specifically, a mention of the Fiat Spider. With the right signals in place, a computer will recognize that the word “Fiat” refers to a car, not some authoritative decree, and that “Spider” is a type of Fiat car, not a creepy bug with eight long legs.

With this semantic information embedded in the text, not only can a human look at the text and appreciate the very small number of people driving a Fiat Spider in Peoria, but people interested in the Fiat Spider car can use computers and search engines to find this text knowing for certain it is about the car and not the bug. If the text signals which Peoria it refers to—the one in Illinois or the one in Arizona—people interested in one or the other city could learn more information more quickly as well. The difference matters: fourteen drivers of the Fiat Spider in Peoria, Illinois, is indeed a low number. Fourteen drivers of that one car in tiny Peoria, Arizona, is a lot.

There are many ways of putting signals into documents—and not only text documents, but also audio and video files—to make them more informative. There is al-

When data is published in machine-readable ways, its meanings can come to life, and it can be the foundation of truly transparent government.

most no end to what can be done with this kind of signaling in webpages or in other documents and data. HTML is a format that it is well known and followed by most Web publishers and browsers across the globe, which is one of the things that makes the Web so powerful and important. Nobody ever has to ask for a more transparent Web page; the use of a widely recognized format takes care of that problem.

Metadata

The term of art for this kind of signaling, done by embedding information in documents or data, is metadata. Metadata is a sort of "who, what, when, and where" that is one step removed from the principal data being collected and presented. It helps a user of the data understand its meanings and importance.

Machine-readability, machine-discoverability, availability, and authoritative sourcing can produce tremendous advances in government transparency.

Here's a familiar example of metadata: lots of peoples' photographs and home videos from the 80s and 90s have a date stamp in the picture, because cameras could be programmed to insert this information into the image (or perhaps it was hard to keep the date stamp out . . .). That metadata allows someone looking at the image later to know when the picture or video was shot. Thus, parents can know the ages of their children in photos, which vacation trip the image is from, and so on. Metadata helps make data more complete and useful.

Metadata can create powerful efficiencies. Say a group of cattle ranchers wants to manage their herds in concert, but maintain separate ownership. They can save money and expense if they all use the same pens and fields, feed their animals together, and so on. Before they move their herds together, they might attach to the ears of each of their cattle a distinctive tag to indicate who is the owner. Then, when the time comes to divide up their herds, this can easily be done.

They can do much more this same way, though. If juvenile animals require different feed than the mature ones, a tag indicating the age of each animal might allow them to be sorted appropriately at feeding time. Another tag might indicate what inoculations each animal has gotten so that disease management of the herd is streamlined. Each of the many "use cases" for managing a herd can be facilitated by metadata that is physically attached to each animal via the ear tag.

The use cases for government data, and thus the metadata needed in government data, are many. Some people will want to see how bills affect existing laws, existing programs, or agencies. Each of these things can be highlighted in documents and discussions so that they are easily found. Some people will want to follow appropriations and spending, so metadata for dollar proposals and dollar-oriented discussions are worthwhile. Other people will want to know what regions, states, localities, parks, buildings, or installations are the subject of documents and debate. And the corporations, associations, and people who take part in public policy processes are of keen interest. All these things—and more—should be in the metadata of government-published information, and the data should be structured so that rich troves of meaningful information are readily apparent in both documents and data. This will make the relevance of documents and information immediately apparent to various interests using computers to scan the information environment. This is machine-readability, and it is the publication practice that will bring government transparency to fruition.

Machine-readability, machine-discoverability, availability, and authoritative sourcing can produce tremendous advances in government transparency. Well-published data about governments' deliberations, management, and results will inform people better and empower them to do a better job of overseeing their governments.

Conclusion

Government transparency is a widely agreed-upon value, but it is agreed upon as a means toward various ends. Libertarians and conservatives support transparency because of their belief that it will expose waste and bloat in government. If the public understands the workings and failings of government better, the demand for government solutions will fall and democracy will produce more libertarian outcomes. American liberals and progressives support transparency because they believe it will validate and strengthen government programs. Transparency will root out corruption and produce better outcomes, winning the public's affection and support for government.

Though the goals may differ, pan-ideological agreement on transparency can remain. Libertarians should not prefer large government programs that are failing. If transparency makes government work better, that is preferable to government working poorly. If the libertarian vision prevails, on the other hand, and transparency produces demand for less government and greater private authority, that will be a result of democratic decisionmaking that all should respect and honor.

The publication practices described here—authoritative sourcing, availability, machine-discoverability, and machine-readability—can help make government more transparent. Governments should publish data about their deliberations, management, and results following these good data practices.

But transparency is not an automatic or instant result of following these good practices, and it is not just the form and formats of data. It turns on the capacity of the society to interact with the data and make use of it. American society will take some time to make use of more transparent data once better practices are in place. There are already thriving communities of researchers, journalists, and software developers using unofficial repositories of government data. If they can do good work with incomplete and imperfect data, they will do even better work with rich, complete data issued promptly by authoritative sources. When fully transparent data comes online, though, researchers will have to learn about these data sources and begin using them. Government transparency and advocacy websites will have to do the same. Government entities themselves will discover new ways to coordinate and organize based on good data-publication practices. Reporters will learn new sources and new habits.

By putting out data that is "liquid" and "pure," governments can meet their responsibility to be transparent, and they can foster this evolution toward a body politic that better consumes data. Transparency is likely to produce a virtuous cycle in which public oversight of government is easier, in which the public has better access to factual information, in which people have less need to rely on ideology, and in which artifice and spin have less effectiveness. The use of good data in some areas will draw demands for more good data in other areas, and many elements of governance and public debate will improve.

Both government and civil society have obligations to fulfill if government transparency is to be a reality. By publishing data optimized for transparency, governments can put the ball back into the court of the transparency advocates.

Transparency turns on the capacity of the society to interact with data and make use of it.

Notes

[1] Barack Obama, "The Change We Need in Washington" (speech, Green Bay, WI, September 22, 2008), http://www.youtube.com/watch?v=o5t8GdxFYBU.

[2] John Boehner Introduces the House GOP Congressional Transparency Initiative, http://www.youtube.com/watch?v=hDr70qRv_9k.

[3] Barack Obama, "Memorandum for the Heads of Executive Departments and Agencies" January 21, 2009, http://www.whitehouse.gov/the_press_office/Transparency_and_Open_Government/.

[4] Peter R. Orszag, "Memorandum for the Heads of Executive Departments and Agencies, Subject: Open Government Directive," M10-06, December 8, 2009, http://www.whitehouse.gov/open/documents/open-government-directive.

[5] The White House, "Open Government Initiative," http://www.whitehouse.gov/open.

6. The White House, "Open Government Initiative Blog," http://www.whitehouse.gov/open/blog.

7. Data.gov, http://www.data.gov/.

8. John Wonderlich, "House Rules Transparency Victory," *The Sunlight Foundation Blog*, December 22, 2010, http://sunlightfoundation.com/blog/2010/12/22/house-rules-transparency-victory/.

9. John Boehner, "Keeping the Pledge: New Majority to Make Legislative Data More Open, Accessible," *John Boehner* (blog), April 29, 2011, http://www.johnboehner.house.gov/Blog/?postid=238790.

10. Orszag.

11. Jim Harper, "Grading Agencies' High-Value Datasets," *Cato@Liberty* (blog), February 5, 2010, http://www.cato-at-liberty.org/grading-agencies-high-value-data-sets/.

12. Cato Institute, "Just Give Us the Data! Prospects for Putting Government Information to Revolutionary New Uses," Policy Forum, December 10, 2008, http://www.cato.org/event.php?eventid=5475.

13. "8 Principles of Open Government Data," December 8, 2007, http://www.opengovdata.org/home/8principles [hereinafter "Open Government Principles"].

14. "Ten Principles for Opening Up Government Information," August 11, 2011, http://sunlightfoundation.com/policy/documents/ten-open-data-principles/.

15. Open House Project, "Congressional Information & the Internet: A Collaborative Examination of the House of Representatives and Internet Technology," May 8, 2007, http://www.theopenhouseproject.com/the-open-house-project-report/.

16. A catalog of many such documents can be found on Open Government Data's website, at http://www.opengovdata.org/home/reading-list.

17. The second Open Government Data Principle called for data "published as collected at the source, with the finest possible level of granularity, not in aggregate or modified forms." Open Government Principles.

18. Government Printing Office, FDsys, http://www.gpo.gov/fdsys/.

19. Library of Congress, THOMAS, http://thomas.loc.gov/home/thomas.php.

20. Govtrack.us, http://www.govtrack.us/.

21. The third Open Government Data Principle is making data "available as quickly as necessary to preserve the value of the data."

22. Author's correspondence, on file.

23. The Open Government Data Principles document called for a contact person "designated to respond to people trying to use the data."

24. The fourth Open Government Data Principle called for accessibility, defined as "available to the widest range of users for the widest range of purposes."

25. This "nearly implicit" principle is featured by Josh Tauberer in his paper, "Open Data is Civic Capital: Best Practices for 'Open Government Data,'" January 29, 2011, http://razor.occams.info/pubdocs/opendataciviccapital.html.

26. The first of the Open Government Data Principles was that data should be "complete." The Association for Computing Machinery recommends: "Citizens should be able to download complete datasets of regulatory, legislative or other information, or appropriately chosen subsets of that information, when it is published by government." Association for Computing Machinery, "ACM U.S. Public Policy Committee (USACM) Recommendations on Open Government," http://www.acm.org/public-policy/open-government.

27. Tauberer.

28. Open Government Data Principles six, seven, and eight address availability. Access must be "non-discriminatory" (available to anyone, with no requirement of registration) (principle 6); nonproprietary (principle 7); and license-free (principle 8).

29. There may be some justified restrictions on availability. Repeated bulk downloads are a form of attack on a data system meant to disable it or render it costly to maintain, for example. This form of attack justifies a gating mechanism on downloads that is entirely reasonable if applied neutrally and carefully.

30. See Wikipedia, "Hypertext Transfer Protocol," http://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol.

31. See Wikipedia, "HTML," http://en.wikipedia.org/wiki/HTML.

32. Jim Harper, "Sunlight Before Signing: Turning the Corner!" *Cato@Liberty* (blog), December 18,

2009, http://www.cato-at-liberty.org/sunlight-before-signing-turning-the-corner/.

33. WashingtonWatch.com, "Earmarks 2011: 39,000+ and $130 Billion," *WashingtonWatch.com* (blog), December 7, 2010), http://www.washingtonwatch.com/blog/2010/12/07/earmarks-2011-39000-and-130-billion/. WashingtonWatch.com is a transparency website run by the author and is unaffiliated with the Cato Institute.

34. *See* Wikipedia, "Internet Media Type," http://en.wikipedia.org/wiki/Internet_media_type.

35. Federal Election Commission, "Electronically Filed Reports and Statements," http://www.fec.gov/finance/disclosure/ftpefile.shtml.

36. The fifth Open Government Data Principle called for machine processability, in which "Data are reasonably structured to allow automated processing of it."

37. Tauberer.

38. Association for Computing Machinery.

39. "Biographical Directory of the United States Congress: 1774–present," http://bioguide.congress.gov/biosearch/biosearch.asp.

40. Tauberer notes: "To the extent two data sets refer to the same kinds of things, the creators of the data sets should strive to make them interoperable. This may mean developing a shared data standard, or adopting an existing standard, possibly through coordination within government across agencies."

Appendix II

Grading the Government's Data Publication Practices

Policy Analysis

No. 711 November 5, 2012

Grading the Government's Data Publication Practices

by Jim Harper

Executive Summary

Barack Obama promised transparency and open government when he campaigned for president in 2008, and he took office aiming to deliver it. Today, the federal government is not transparent, and government transparency has not improved materially since the beginning of President Obama's administration. This is not due to lack of interest or effort, though. Along with meeting political forces greater than his promises, the Obama transparency tailspin was a product of failure to apprehend what transparency is and how it is produced.

A variety of good data publication practices can help produce government transparency: authoritative sourcing, availability, machine-discoverability, and machine-readability. The Cato Institute has modeled what data the government should publish in the areas of legislative process and budgeting, spending, and appropriating. The administration and the Congress both receive fairly low marks under systematic examination of their data publication practices.

Between the Obama administration and House Republicans, the former, starting from a low transparency baseline, made extravagant promises and put significant effort into the project of government transparency. It has not been a success. House Republicans, who manage a far smaller segment of the government, started from a higher transparency baseline, made modest promises, and have taken limited steps to execute on those promises. President Obama lags behind House Republicans, but both have a long way to go.

Jim Harper is director of information policy studies at the Cato Institute and the webmaster of WashingtonWatch.com.

CATO INSTITUTE

There was no lack of effort or creativity around data transparency at the outset of the Obama Administration.

Introduction

As a campaigner in 2008, President Obama promised voters hope, change, and transparency.[1] Within minutes of his taking office on January 20, 2009, in fact, the Whitehouse.gov website declared: "President Obama has committed to making his administration the most open and transparent in history."[2] His first presidential memorandum, issued the next day, was entitled "Transparency and Open Government." It declared:

> My Administration is committed to creating an unprecedented level of openness in Government. We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration. Openness will strengthen our democracy and promote efficiency and effectiveness in Government.[3]

The road to government transparency is long. Nearly four years later, few would argue that American democracy has materially strengthened, or that the government is any more effective and efficient, due to forward strides in transparency and openness. Indeed, the administration has come under fire recently—as every administration does, it seems—for significant transparency failings.

Freedom of Information Act (FOIA) policy is an example. In its early days, the Obama administration committed to improving the government's FOIA practices. In March 2009 Attorney General Eric Holder issued a widely lauded memorandum ordering improvements in FOIA compliance.[4] But this September, Bloomberg news reported on its test of the Obama Administration's commitment to transparency under FOIA. Bloomberg found that 19 of 20 cabinet-level agencies disobeyed the public disclosure law when it asked for information about the cost of agency leaders' travel. Just 8 of 57 federal agencies met Bloomberg's request for documents within the 20-day disclosure window required by the act.[5]

President Obama's campaign promise to post laws to the White House website for five days of public comment before he signed them went virtually ignored by the White House in the first year of his administration. Only recently has he reached two-thirds compliance with the "Sunlight Before Signing" promise, and this is because of the multitude of bills Congress passes to rename post offices and such. More important bills are often given less than the promised five days' sunlight.[6]

There was no lack of effort or creativity around data transparency at the outset of the Obama Administration. In May 2009 White House officials announced on the new *Open Government Initiative* blog that they would elicit the public's input into the formulation of its transparency policies. In a meta-transparency flourish, the public was invited to join in with the brainstorming, discussion, and drafting of the government's policies.[7]

The conspicuously transparent, participatory, and collaborative process contributed something, evidently, to an "Open Government Directive," issued in December 2009 by Office of Management and Budget head Peter Orszag.[8] Its clear focus was to give the public access to data. The directive ordered agencies to publish within 45 days at least three previously unavailable "high-value data sets" online in an open format and to register them with the federal government's data portal, Data.gov. Each agency was to create an "Open Government Webpage" as a gateway to agency activities related to the Open Government Directive.

Many, many of President Obama's transparency promises went by the wayside. His guarantee that health care legislation would be negotiated "around a big table" and televised on C-SPAN was quite nearly the opposite of what occurred.[9] People are free to observe whether it is political immaturity, idealism, or dishonesty that prompted transparency promises of this kind. Whatever the case, history may show that the "high-value data set" challenge was where the Obama Administration's data transparency effort began its tailspin.

Celebrated though it is, transparency is not a well-defined concept, and the administration's most concerted effort to deliver it missed the mark. The reason is that the definition of "high-value data set" it adopted was hopelessly vague:

> High-value information is information that can be used to increase agency accountability and responsiveness; improve public knowledge of the agency and its operations; further the core mission of the agency; create economic opportunity; or respond to need and demand as identified through public consultation.

Essentially anything agencies wanted to publish they could publish claiming "high value" for it.

Agencies "adopted a passive-aggressive attitude" toward the Data.gov effort, according to political scientist Alon Peled.[10] They technically complied with the requirements of the Open Government Memorandum, but did not select data that the public valued.

The Open Government Directive allowed agencies to exploit a subtle "shift in vocabulary" in the area of open government. They diverted the project away from the core government transparency that the public found so attractive about President Obama's campaign claims. "The term 'open government data' might refer to data that makes the government as a whole more open (that is, more publicly accountable)," write Harlan Yu and David Robinson, "or instead might refer to politically neutral public sector disclosures that are easy to reuse, even if they have nothing to do with public accountability."[11]

The Agriculture department published data about the race, ethnicity, and gender of farm operators, for example, rather than about the funds it spent to collect that kind of information. An informal Cato Institute study examining agencies' "high-value" data feeds found, "almost uniformly, the agencies came up with interesting data—but 'interesting' is in the eye of the beholder. And interesting data collected by an agency doesn't necessarily give the insight into government we were looking for."[12]

Genuinely high-value data for purposes of government transparency would provide insight in three areas not found in many of the early Data.gov feeds. True high-value data would be about government entities' management, deliberations, or results.[13]

"Open data can be a powerful force for public accountability," write Yu and Robinson, "It can make existing information easier to analyze, process, and combine than ever before, allowing a new level of public scrutiny."[14] This is undoubtedly true, and Americans have experienced vastly increased access to information in so many walks of life—shopping, news-gathering, and investments, to name just three. Data-starved public oversight of government appears sorely lacking in comparison.

In September a new transparency-related international initiative took center stage for the administration, the Open Government Partnership (OGP).[15] This "multilateral initiative" was created "to promote transparency, fight corruption, strengthen accountability, and empower citizens."[16] Participating countries pledged "to undertake meaningful new steps as part of a concrete action plan, developed and implemented in close consultation with their citizens." The OGP website touts a panoply of meetings, plans, and social media outreach efforts, and a recent graphic displayed on the home page said in bold letters, "From Commitment to Action." Its authors probably have no sense of the irony in that declaration. Significant actions, after all, announce themselves.

Nothing about the OGP is harmful, and it may produce genuine gains for openness in participating countries. However, it has not produced, and does not hold out, the fundamental change—data-oriented change—that

Celebrated though it is, transparency is not a well-defined concept.

was at the heart of President Obama's campaign promises.

The Obama administration is not the only actor on the federal stage, of course. House Republicans made transparency promises of their own in the course of their campaign to retake control of the House of Representatives, which they did in 2011.

"The lack of transparency in Congress has been a problem for generations, under majorities Republican and Democrat alike," said aspiring House speaker John Boehner (R-OH) in late 2009. "But with the advent of the Internet, it's time for this to change."[17]

Since 1995, the Library of Congress's THOMAS website has published information, sometimes in the form of useful data, about Congress and its activities. Upon taking control of the House for the first time in 40 years, the Republican leadership of the 104th Congress directed the Library of Congress to make federal legislative information freely available to the public. The offerings on the site now include bills, resolutions, activity in Congress, the *Congressional Record*, schedules, calendars, committee information, the president's nominations, and treaties.[18]

In an attempt to improve the availability of key information, at the beginning of the 112th Congress the House instituted a rule—not always complied with—that bills should be posted online for three calendar days before receiving a vote on the House floor.[19] The House followed up by creating a site at data.house.gov where such bills are posted. In February 2012 the House Committee on Administration held a day-long conference on legislative data,[20] evidence of continuing interest and of plans to move forward. And in September, the Library of Congress debuted beta.congress.gov, which is slated to be the repository for legislative data that ultimately replaces the THOMAS website.[21]

The transparency problem is far from solved.

Between the Obama administration and House Republicans, the former, starting from a low transparency baseline, made extravagant promises and put significant effort into the project of government transparency. It has not been a success. House Republicans, who manage a far smaller segment of the government, started from a higher transparency baseline, made modest promises, and have taken limited steps to execute those promises.

The transparency problem is far from solved, of course. The information that the public would use to increase their oversight and participation is still largely inaccessible. The Republican House may be ahead, but both the administration and Congress score poorly under systematic examination of their data publication practices.

The Data that Would Make for Transparent Government

It was not disinterest that caused the Obama administration transparency effort to fade. Arguably, it was the failure of the transparency community to ask clearly for what it wants: good data about the deliberations, management, and results of government entities and agencies. So in January 2011 the Cato Institute began working with a wide variety of groups and advisers to "model" governmental processes as data and then to prescribe how this data should be published.

Data modeling is arcane stuff, but it is worth understanding here at the dawn of the Information Age. "Data" is collected abstract representations of things in the world. We use the number "3," for example, to reduce a quantity of things to an abstract, useful form—an item of data. Because clerks can use numbers to list the quantities of fruits and vegetables on hand, store managers can effectively carry out their purchasing, pricing, and selling instead of spending all of their time checking for themselves how much of everything there is. Data makes everything in life a little easier and more efficient for everyone.

Legislative and budgetary processes are not a grocery store's produce department, of course. They are complex activities involving many actors, organizations, and steps. The Cato Institute's modeling of these processes reduced everything to "entities," each having various "properties." The entities and their properties describe the things in legislative and budgetary processes and the logical relationships among them, like members of Congress, the bills they introduce, hearings on the bills, amendments, votes, and so on. The "entity" and "property" terminology corresponds with usage in the world of data management, it is used to make coding easier for people in that field, and it helps to resolve ambiguities in translating governmental processes into useful data. The modeling was restricted to formal parts of the processes, excluding, for example, the varied organizations that try to exert influence, informal communications among members of Congress, and so on.

The project also loosely defined several "markup types," guides for how documents that come out of the legislative process should be structured and published to maximize their utility. The models and markup types are discussed in a pair of *Cato@Liberty* blog posts that also issued preliminary grades on the quality of data publication about the entities.[22] The models and markup types for legislative data and budgeting/appropriations/spending data can be found in Appendixes A and B, respectively.

Next, the project examined the publication methods that allow data to reach its highest and best use. Four key data practices that support government transparency emerged. Documented in a Cato Institute Briefing Paper entitled "Publication Practices for Transparent Government,"[23] those practices are authoritative sourcing, availability, machine-discoverability, and machine-readability.

Authoritative sourcing means producing data as near to its original source and time as possible, so that the public uniformly comes to rely on the best sources of data. The second transparent data practice, availability, entails consistency and confidence in data, including permanence, completeness, and good updating practices.

The third transparent data practice, machine-discoverability, occurs when information is arranged so that a computer can discover the data and follow linkages among it. Machine-discoverability exists when data is presented consistently with a host of customs about how data is identified and referenced, the naming of documents and files, the protocols for communicating data, and the organization of data within files.

The fourth transparent data practice, machine-readability, is the heart of transparency because it allows the many meanings of data to be discovered. Machine-readable data is logically structured so that computers can automatically generate the myriad stories that the data has to tell and put it to the hundreds of uses the public would make of it in government oversight. A common and popular language for structuring and containing data is called XML, or eXtensible Markup Language, which is a relative of HTML (hypertext markup language), the language that underlies the World Wide Web.

Beginning in September 2011 the project graded how well Congress and the administration publish data about the key entities in the processes they oversee. Congress is responsible for data pertaining to the legislative process, of course. The administration has the bulk of the responsibility for budget-related data (except for the congressional budgets and appropriations). These grades are available in a pair of *Cato@Liberty* blog posts[24] and in Appendixes C and D.

With the experience of the past year, the project returned to grading in September 2012. With input from staff at GovTrack.us, the National Priorities Project, OMB Watch, and the Sunlight Foundation (their endorsement of the grades not implied by their assistance), we assessed how well data is now published. The grades presented in Figures 1 and 2 are largely consistent with the prior year—little changed between the two grading periods—but there were some changes in grades in both directions due to improvements in publication, discovery of data sources by our panel of graders, and

Four key data practices support government transparency: authoritative sourcing, availability, machine-discoverability, and machine-readability.

heightened expectations. “Incompletes” given in the first year of grading became Fs in some cases and Ds in others.

It is important to highlight that grades are a lagging indicator. Transparency is not just a product of good data publication, but also of the society’s ability to digest and use information. Once data feeds are published, it takes a little while for the community of users to find them and make use of them. A new web site dedicated to congressional information, beta.congress.gov, will undoubtedly improve data transparency and the grades for data it publishes, assuming it lives up to expectations.

Government transparency is a widely agreed-upon value, sought after as a means toward various ends.

Government transparency is a widely agreed-upon value, sought after as a means toward various ends. Libertarians and conservatives support transparency because of their belief that it will expose waste and bloat in government. If the public understands the workings and failings of government better, the demand for government solutions will fall and democracy will produce more libertarian outcomes. American liberals and progressives support transparency because they believe it will validate and strengthen government programs. Transparency will root out corruption and produce better outcomes, winning the public’s affection and support for government.

Though the goals may differ, pan-ideological agreement on transparency can remain. Libertarians should not prefer large government programs that are failing. If transparency makes government work better, that is preferable to government working poorly. If the libertarian vision prevails, on the other hand, and transparency produces demand for less government and greater private authority, that will be a result of democratic decisionmaking that liberals and progressives should respect and honor.

With that, here are the major entities in the legislative process and in budgeting, appropriating, and spending; the grades that reflect the quality of the data published about them; and a discussion of both.

Publication Practices for Transparent Government: Rating Congress

House Membership: C-
Senate Membership: A-

It would seem simple enough to publish data about who holds office in the House of Representatives and Senate, and it is. There are problems with the way the data is published, though, which the House and Senate could easily remedy.

On the positive side—and this is not to be discounted—there is a thing called the “Biographical Directory of the United States Congress,” a compendium of information about all present and former members of the U.S. Congress (as well as the Continental Congress), including delegates and resident commissioners. The “Bioguide” website at bioguide.congress.gov is a great resource for searching out historical information.

But there is little sign that Bioguide is Congress’s repository of record, and it is little known by users, giving it lower authority marks than it should have. Some look to the House and Senate websites and beta.congress.gov for information about federal representatives, splitting authority among websites, rather than one established and agreed upon resource.

Bioguide scores highly on availability—we know of no problems with up-time or completeness (though it could use quicker updating when new members are elected). Bioguide is not structured for discoverability, though. Most people have not seen it, because search engines are not finding it.

Bioguide does a good thing in terms of machine readability, though. It assigns a unique ID to each of the people in its database. This is the first, basic step in making data useful for computers, and the Bioguide ID should probably be the standard for machine identification of elected officials wherever they are referred to in data. Unfortunately, the biographical content in Bioguide is not machine-readable.

Figure 1

October 2012

Publication Practices for Transparent Government: Rating the Congress

How well can the Internet access data about Congress' work? The Cato Institute rated how well Congress publishes information in terms of authoritative sourcing, availability, machine-discoverability, and machine-readability.

Subject	Grade	Comments
House and Senate Membership	House C- Senate A-	*The Senate has taken the lead on making data about who represents Americans in Washington machine-readable.*
Committees and Subcommittees	C-	*Organizing and centralizing committee information would create a lot of clarity with a minimum of effort.*
Meetings of House, Senate, and Committees	House B Senate B	*The House has improved its data about floor debates. The Senate is strong on commitee meetings.*
Meeting Records	D-	*There is lots of work to do before transcripts and other meeting records can be called transparent.*
Committee Reports	C+	*Committee reports can be found, but they're not machine-readable.*
Bills	B-	*Bills are the "pretty-good-news" story in legislative transparency, though there is room for improvement.*
Amendments	F	*Amendments are hard to track in any systematic way—and Congress has done little to make them trackable.*
Motions	F	*If the public is going to have insight into the decisions Congress makes, the motions on which Congress acts should be published as data.*
Decisions and Votes	B+	*Vote information is in good shape, but voice votes and unanimous consents should be published as data.*
Communications (Inter- and Intra-Branch)	F	*Transparent access to the messages sent among the House, Senate, and executive branch would complete the picture available to the public.*

Figure 2

Publication Practices for Transparent Government: Budgeting, Appropriating, and Spending

How well can the Internet access data about the federal government's budgeting, appropriating, and spending? The Cato Institute rated how well the government publishes information in terms of authoritative sourcing, availability, machine-discoverability, and machine-readability.

Subject	Grade	Comments
Agencies	D-	*This grade is generous. There really should be a machine-readable federal government "organization chart."*
Bureaus	D-	*The sub-units of agencies have the same problem.*
Programs	D	*Program information is obscure, incomplete, and unorganized.*
Projects	F	*Some project information gets published, but the organization of it is bad.*
Budget Documents	Congress D White House B-	*The president's budget submission and congressional budget resolutions are a mixed bag.*
Budget Authority	F	*Legal authority to spend is hidden and unstructured.*
Warrants, Apportionments, and Allocations	F	*Spending authority is divided up in an opaque way.*
Obligations	B-	*Commitments to spend taxpayer money are visible some places.*
Parties	F	*A proprietary identifier system makes it hard to know where the money is going.*
Outlays	C-	*We need real-time, granular spending data.*

As noted above, the other ways of learning about House and Senate membership are ad hoc. The Government Printing Office has a "Guide to House and Senate Members" at http://memberguide.gpo.gov/ that duplicates information found elsewhere. The House website presents a list of members along with district information, party affiliation, and so on, in HTML format (http://www.house.gov/representatives/), and beta.congress.gov does as well (http://beta.congress.gov/members/). Someone who wants a complete dataset must collect data from these sources using a computer program to scrape the data and through manual curation. The HTML presentations do not break out key information in ways useful for computers. The Senate membership page,[25] on the other hand, includes a link to an XML representation that is machine readable. That is the reason why the Senate scores so well compared to the House.

Much more information about our representatives flows to the public via representatives' individual websites. These are nonauthoritative websites that search engine spidering combines to use as a record of the Congress's membership. They are available and discoverable, again because of that prime house.gov and senate.gov real estate. But they only reveal data about the membership of Congress incidentally to communicating the press releases, photos, and announcements that representatives want to have online.

It is a narrow point, but there should be one and only one authoritative, well-published source of information about House and Senate membership from which all others flow. The variety of sources that exist combine to give Congress pretty good grades on publishing information about who represents Americans in Washington, but improving in this area is a simple matter of coordinated House and Senate efforts.

There should be one and only one authoritative, well-published source of information about House and Senate membership.

Committees and Subcommittees: C-

Like Americans' representation in Congress, lists of committees, their membership, and jurisdiction should be an easy lift. But it is not as easy as it should be to learn about the committees to which Congress delegates much of its work and the subcommittees to which the work gets further distributed.

The Senate has committee names and URLs prominently available on its main website.[26] The House does, too, at http://house.gov/committees/. But neither page offers machine-readable information about committees and committee assignments. The Senate has a nice list of committee assignments, again, though, not machine-readable. The House requires visitors to click through to each committee's web page to research what they do and who serves on them. For that, you'd go to individual committee websites, each one different from the others. There is an authoritative list of House committees with unique identifiers,[27] but it's published as a PDF, and it is not clear that it is used elsewhere for referring to committees.

Without a recognized place to go to get data about committees, this area suffers from lacking authority. To the extent there are data, availability is not a problem, but machine-discoverability suffers for having each committee publish distinctly, in formats like HTML, who their members are, who their leaders are, and what their jurisdiction is.

With the data scattered about this way, the Internet can't really see it. More prominence, including data such as subcommittees and jurisdiction, and use of a recognized set of standard identifiers would take this resource a long way.

Until committee data are centrally published using standard identifiers (for both committees and their members), machine-readability will be very low. The Internet makes sense of congressional committees as best it can, but a whole lot of organizing and centralizing—with a definitive, always-current, and machine-readable record of committees, their memberships, and their jurisdictions—would create a lot of clarity in this area with a minimum of effort.

Can the public learn easily about what meetings are happening, where they are happening, when they are happening, and what they are about? It depends on which side of the Capitol you're on.

Meetings of House, Senate, and Committees—House: B/Senate: B

When the House, the Senate, committees, and subcommittees have their meetings, the business of the people is being done. Can the public learn easily about what meetings are happening, where they are happening, when they are happening, and what they are about? It depends on which side of the Capitol you're on.

The Senate is pretty good about publishing notices of committee meetings. From a webpage with meeting notices listed on it,[28] there is a link to an XML version of the data to automatically inform the public.

If a particular issue is under consideration in a Senate committee meeting, this is a way for the public to learn about it. This is authoritative, it is available, it is machine-discoverable, and has some machine-readable features. That means any application, website, researcher, or reporter can quickly use these data to generate more—and more useful—information about Congress.

The House does not have anything similar for committee meetings. To learn about those meetings, one has to scroll through page after page of committee announcements or calendars. Insiders subscribe to paid services. The House can catch up with the Senate in this area.

Where the House excels and the Senate lags is in notice about what will be considered on the floor. The House made great strides with the institution of docs.house.gov, which displays legislation heading for the floor. This allows any visitor, and various websites and services, to focus their attention on the nation's business for the week.

Credit is due the House for establishing this resource and using it to inform the public using authoritative, available, and machine-discoverable and -readable data. This is an area where the Senate has the catching up to do.

For different reasons, the House and Senate both garner Bs. Were they to copy the best of each other, they would both have As.

Meeting Records: D-

There is a lot of work to do before meeting records can be called transparent. The *Congressional Record* is the authoritative record of what transpires on the House and Senate floors, but nothing similar reveals the content of committee meetings. Those meeting records are produced after much delay—sometimes an incredibly long delay—by the committees themselves. These records are obscure, and they are not being published in ways that make things easy for computers to find and comprehend.

In addition, the *Congressional Record* doesn't have the machine-discoverable publication or machine-readable structure that it could and should. Giving unique, consistent IDs in the *Record* to members of Congress, to bills, and other regular subjects of this publication would go a long way to improving it. The same would improve transcripts of committee meetings.

Another form of meeting record exists: videos. These have yet to be standardized, organized, and published in a reliable and uniform way, but the HouseLive site (http://houselive.gov/) is a significant step in the right direction. It will be of greater use when it can integrate with other records of Congress. Real-time flagging of members and key subjects of debate in the video stream would be a great improvement in transparency. Setting video and video meta-data standards for use by both Houses of Congress, by committees, and by subcommittees would improve things dramatically.

House video is a bright spot in a very dark field, but both will shine brighter in time. When the surrounding information environment has improved to educate the public about goings-on in Congress in real time, the demand for and usefulness of video will increase.

Committee Reports: C+

Committee reports are important parts of the legislative process, documenting the findings and recommendations that committees report to the full House and Senate. They do see publication on the most authoritative resource for committee reports, the Library of Congress's THOMAS system. They are also published by the Government Printing Office.[29] The GPO's Federal Digital System (FDsys) is relatively new and is meant to improve systematic access to government documents, but it has not become recognized as an authoritative source for many of those documents.

Because of the sources through which they are published, committee reports are somewhat machine-discoverable, but without good semantic information embedded in them, committee reports are barely visible to the Internet.

Rather than publication in HTML and PDF, committee reports should be published fully marked up with the array of signals that reveal what bills, statutes, and agencies they deal with, as well as authorizations and appropriations, so that the Internet can discover and make use of these documents.

Bills: B-

Bills are a "pretty-good-news" story in legislative transparency. Most are promptly published. It would be better, of course, if they were all immediately published at the moment they were introduced, and if both the House and Senate published last-minute, omnibus bills before debating and voting on them.

A small gap in authority exists around bills. Some people look to the Library of Congress and the THOMAS site, and now beta.congress.gov, for bill information. Others look to the Government Printing Office. Which is the authority for bill content? This issue has not caused many problems so far. Once published, bill information remains available, which is good.

Publication of bills in HTML on the THOMAS site makes them reasonably machine-discoverable. Witness the fact that searching for a bill will often turn up the version at that source.

Where bills could improve some is in their machine-readability. Some information such as sponsorship and U.S. code references is present in the bills that are published in XML, and nearly all bills are now published in XML, which is great. Much more information should be published machine-readably in bills, though, such as references to agencies and programs, to states or localities, to authorizations and appropriations, and so on, referred to using standard identifiers.

With the work that the THOMAS system does to gather information in one place, bill data are good. This is relative to other, less-well-published data, though. There is yet room for improvement.

Bills are a "pretty-good-news" story in legislative transparency.

Amendments: F

Amendments are not the good-news story that bills are. They are "barely available," says Eric Mill of the Sunlight Foundation. "Given that amendments (especially in the Senate) can be as large and important as original legislation, this is an egregious oversight."

With a few exceptions, amendments are hard to track in any systematic way. When bills come to the House and Senate floors, amendment text is often available, but amendments are often plopped somewhere in the middle of the *Congressional Record* without any reliable, understood, machine-readable connection to the underlying legislation. It is very hard to see how amendments affect the bills they would change.

In committees, the story is quite a bit worse. Committee amendments are almost completely opaque. There is almost no publication of amendments at all—certainly not amendments that have been withdrawn or defeated. Some major revisions in process are due if committee amendments are going to see the light of day as they should.

Motions: F

When the House, the Senate, or a committee is going to take some kind of action, it does so on the basis of a motion. If the public is going to have insight into the decisions Congress makes, it should have access to the motions on which Congress acts.

But motions are something of a black hole. Many of them can be found in the *Congressional Record*, but it takes a human who understands legislative procedure and who is willing to read the *Congressional Record* to find them. That is not modern transparency.

Motions can be articulated as data. There are distinct types of motions. Congress can publish which meeting a motion occurs in, when the motion occurs, what the proposition is, what the object of the motion is, and so on. Along with decisions, motions are key elements of the legislative process. They can and should be published as data.

Decisions and Votes: B+

When a motion is pending, a body such as the House, the Senate, or a committee will make a decision on it, only sometimes using votes. These decisions are crucial moments in the legislative process, which should be published as data. Like motions, many decisions are not yet published usefully. Decisions made without a vote in the House or Senate are published in text form as part of the *Congressional Record*, but they are not published as data, so they remain opaque to the Internet. Many, many decisions come in the form of voice votes, unanimous consents, and so on.

Voting puts members of Congress on record about where they stand. And happily, vote information is in pretty good shape. Each chamber publishes data about votes, meaning authority is well handled. Vote data are available and timely.

Voting puts members of Congress on record about where they stand. And happily, vote information is in pretty good shape.

Both the House[30] and Senate[31] produce vote information. The latter also publishes roll call tables in XML, which is useful for computer-aided oversight. Overall, voting data are pretty well handled. But the omission of voice votes and unanimous consents drags the grade down and will drag it down further as the quality of data publication in other areas rises.

Communications (Inter- and Intra-Branch): F

The Constitution requires each house of Congress to "keep a Journal of its Proceedings, and from time to time publish the same." The basic steps in the legislative process (discussed elsewhere) go into the journals of the House and Senate, along with communications among governmental bodies.

These messages, sent among the House, Senate, and Executive Branch, are essential parts of the legislative process, but they do not see publication. Putting these communications online—including unique identifiers, the sending and receiving body, any meeting that produced the communication, the text of the communication, and key subjects such as bills—would complete the picture that is available to the public.

Publication Practices for Transparent Government: Budgeting, Appropriations, and Spending

Agencies: D-

Federal agencies are the "agents" of Congress and the president. They carry out federal policy and spending decisions. Accordingly, one of the building blocks of data about spending is going to be a definitive list of the organizational units that do the spending.

Is there such a list? Yes. It's Appendix C of OMB Circular A-11, entitled: "Listing of OMB Agency/Bureau and Treasury Codes." This is a poorly organized PDF document that is found on the Office of Management and Budget website.[32]

Poorly organized PDFs are not good transparency. Believe it or not, there is still no federal government "organization chart" that is published in a way amenable to computer processing.

There are almost certainly sets of distinct identifiers for agencies that both the Treasury department and the Office of Management and Budget use. With modifications, either of these could be published as the executive branch's definitive list of its agencies. But nobody has done that. Nobody seems yet to have thought of publishing data about the basic units of the executive branch online in a machine-discoverable and machine-readable format.

In our preliminary grading, we gave this category an "incomplete" rather than an F. That was "beyond generous," according to Becky Sweger of the National Priorities Project. We expect improvement in publication of this data, and the grades will be low until we get it.

Bureaus: D-

The sub-units of agencies are bureaus, and the situation with agencies applies to data about the offices where the work of agencies get divided up. Bureaus have identifiers. It's just that nobody publishes a list of bureaus, their parent agencies, and other key information for the Internet-connected public to use in coordinating its oversight.

Again, a prior "incomplete" in this area has converted to a D-, saved from being an F only by the fact that there is a list, however poorly organized and published, by the Office of Management and Budget.

Programs: D

It is damning with faint praise to call "programs" the brightest light on the organizational-data Christmas tree. The work of the government is parceled out for actual execution in programs. Like information about their parental units, the agencies and bureaus, data that identifies and distinguishes programs is not comprehensively published.

Some information about programs is available in usable form. The Catalog of Federal Domestic Assistance website (www.cfda.gov) has useful aggregation of some information on programs, but the canonical guide to government programs, along with the bureaus and agencies that run them, does not exist.

Programs will be a little bit heavier a lift than agencies and bureaus—the number of programs exceeds the number of bureaus by something like an order of magnitude, much as the number of bureaus exceeds the number of agencies. And it might be that some programs have more than one agency/bureau parent. But today's powerful computers can keep track of these things—they can count pretty high. The government should figure out all the programs it has, keep that list up to date, and publish it for public consumption.

Thanks to the CFDA, data publication about the federal government's programs gets a D.

Projects: F

Projects are where the rubber hits the road. These are the organizational vehicles the government uses to enter into contracts and create other obligations that deliver on government services. Some project information gets published, but the publication is so bad that we give this area a low grade indeed.

Information about projects can be found. You can search for projects by name on USASpending.gov, and descriptions of projects appear in USASpending/FAADS downloads, ("FAADS" is the Federal Assistance Award Data System), but there is no canonical list of projects that we could find. There should be, and there should have been for a long time now.

The generosity and patience we showed in earlier grading with respect to agencies, budgets, and programs has run out. There's more than nothing here, but projects, so essential to have complete information about, gets an F.

**Budget Documents—
Congress: D/White House: B-**

The president's annual budget submission and the congressional budget resolutions are the planning documents that the president and Congress use to map the direction of government spending each year. These documents are published authoritatively, and they are consistently available, which is good. They are sometimes machine-

Believe it or not, there is still no federal government "organization chart" published in a way amenable to computer processing.

Ideally, there would be a nice, neat connection from budget authority right down to every outlay of funds.

discoverable, but they are not terribly machine-readable.

The appendices to the president's budget are published in XML format, which vastly reduces the time it takes to work with the data in them. That's really good. But the congressional budget resolutions—when they exist—have no similar organization, and there is low correspondence between the budget resolutions that Congress puts out and the budget the president puts out. You would think that a person—or better yet, a computer—should be able to lay these documents side by side for comparison, but nobody can.

For its use of XML, the White House gets a B-. Congress gets a D.

Budget Authority: F

"Budget authority" is a term of art for what probably should be called "spending authority." It's the power to spend money, created when Congress and the president pass a law containing such authority.

Proposed budget authority is pretty darn opaque. The bills in Congress that contain budget authority are consistently published online—that's good—but they don't highlight budget authority in machine-readable ways. No computer can figure out how much budget authority is out there in pending legislation.

Existing budget authority is pretty well documented in the Treasury Department's FAST book (Federal Account Symbols and Titles). This handy resource lists Treasury accounts and the statutes and laws that provide their budget authority. The FAST book is not terrible, but the only form we've found it in is PDF. PDF is terrible. And nobody among our graders uses the FAST book.

Congress can do a lot better, by highlighting budget authority in bills in a machine-readable way. The administration can do much, much better than publishing the obscure FAST book in PDF.

Ideally, there would be a nice, neat connection from budget authority right down to every outlay of funds, and back up again from every outlay to its budget authority. These connections, published online in useful ways, would allow public oversight to blossom. But the seeds have yet to be planted.

Warrants, Apportionments, and Allocations: F

After Congress and the president create budget authority, that authority gets divvied up to different agencies, bureaus, programs, and projects. How well documented are these processes? Not well.

An appropriation warrant is an assignment of funds by the Treasury to a treasury account to serve a particular budget authority. It's the indication that there is money in an account for an agency to obligate and then spend. "OMB has a web portal that agencies used to send apportionment requests," notes the National Priorities Project's Becky Sweger, "so the apportionment data are out there."

Where is this warrant data? We can't find it. Given Treasury's thoroughness, it probably exists, but it's just not out there for public consumption.

An apportionment is an instruction from the Office of Management and Budget to an agency about how much it may spend from a Treasury account in service of given budget authority in a given period of time.

We haven't seen any data about this, and we're not sure that there is any. There should be. And we should get to see it.

An allocation is a similar division of budget authority by an agency into programs or projects. We don't see any data on this either. And we should.

These essential elements of government spending should be published for all to see. They are not published, garnering the executive branch an F.

Obligations: B-

Obligations are the commitments to spend money into which government agencies enter. Things like contracts to buy pens, hiring of people to write with those pens, and much, much more.

USASpending.gov has quickly become the authoritative source for this information, but it is not the entire view of spending, and the data is "dirty": inconsistent and unreliable. The use of proprietary DUNS numbers—the Data Universal Numbering System of the firm Dun & Bradstreet—also weakens the availability of obligation data.

There is some good data about obligations, but it is not clean, complete, and well documented. The ideal is to have one source of obligation data that includes every agency, bureau, program, and project. With a decent amount of data out there, though, useful for experts, this category gets a B-.

Parties: F

When the government spends taxpayer dollars, to what parties is it sending the money?

Right now, reporting on parties is dominated by the DUNS number. It provides a unique identifier for each business entity and was developed by Dun & Bradstreet in the 1960s. It's very nice to have a distinct identifier for every entity doing business with the government, but it is not very nice to have the numbering system be a proprietary one.

"Parties" would grade well in terms of machine-readability, which is one of the most important measures of transparency, but because it scores so low on availability, its machine-readability is kind of moot. Until the government moves to an open identifier system for recipients of funds, it will get weak grades on publication of this essential data.

Outlays: C-

For a lot of folks, the big kahuna is knowing where the money goes: outlays. An outlay—literally, the laying out of funds—satisfies an obligation. It's the movement of money from the U.S. Treasury to the outside world.

Outlay numbers are fairly well reported after the fact and in the aggregate. All one has to do is look at the appendices to the president's budget to see how much money has been spent in the past.

But outlay data can be much, much more detailed and timely than that. Each outlay goes to a particular party. Each outlay is done on a particular project or program at the behest of a particular bureau and agency. And each outlay occurs because of a particular budget authority. Right now these details about outlays are nowhere to be found.

"Surely the act of cutting a check doesn't sever all relationship between that amount of money and its corresponding obligation/project/program," writes a frustrated Becky Sweger from the National Priorities Project. "Surely these relationships are intact somewhere and can be published."

Plenty of people inside the government who are familiar with the movement of taxpayer money will be inclined to say, "it's more complicated than that," and it is! But it's going to have to get quite a bit less complicated before these processes can be called transparent.

The time to de-complicate outlays is now. It's a feat of generosity to give this area a C-. That's simply because there is an authoritative source for aggregate past outlay data. As the grades in other areas come up, outlay data that stays the same could go down. Way down.

Conclusion

Many of the entities discussed here are low-hanging fruit if Congress and the administration want to advance transparency and their transparency grades. Authoritative, complete, and well-published lists of House and Senate membership, committees, and subcommittees are easy to produce and maintain, and much of the work has already been done.

The same is true of agencies and bureaus, at least on the executive branch side. Presidential leadership could produce an authoritative list of programs and projects within months. Establishing authoritative identi-

Outlay data can be much, much more detailed and timely.

fiers for these basic units of government is like creating a language, a simple but important language computers can use to assist Americans in their oversight of the federal government.

The more difficult tasks—amendments to legislation, for example, and discretely identified budget authorities—will take some work. But such work can produce massive strides forward in accountable, efficient, responsive, and—in the libertarian vision—smaller government.

Notes

1. An illustration of the focus on the topic, Politifact.com's "Obameter" lists 35 Obama campaign promises from 2008 related to transparency, http://www.politifact.com/truth-o-meter/promises/obameter/subjects/transparency/.

2. Macon Phillips, "Change Has Come to Whitehouse.gov," *The White House Blog*, January 20, 2009 (12:01 p.m. EDT), http://www.whitehouse.gov/blog/change_has_come_to_whitehouse-gov.

3. Barack Obama, "Transparency and Open Government," Presidential Memorandum (January 21, 2012), http://www.whitehouse.gov/the-press-office/transparency-and-open-government.

4. Eric Holder, "Memorandum for Heads of Executive Departments and Agencies: The Freedom of Information Act (FOIA)," Office of the Attorney General (March 19 2009), http://www.justice.gov/ag/foia-memo-march2009.pdf.

5. Jim Snyder and Danielle Ivory, "Obama Cabinet Flunks Disclosure Test with 19 in 20 Ignoring Law," Bloomberg.com, September 28, 2012, http://www.bloomberg.com/news/2012-09-28/obama-cabinet-flunks-disclosure-test-with-19-in-20-ignoring-law.html.

6. Jim Harper, "Sunlight before Signing: Measuring a Campaign Promise," Cato@Liberty (blog), September 6, 2012, http://www.cato-at-liberty.org/sunlight-before-signing-measuring-a-campaign-promise/.

7. Jesse Lee, "Transparency and Open Government," May 21, 2009, http://www.whitehouse.gov/blog/2009/05/21/transparency-and-open-government.

8. Peter R. Orszag, "Memorandum for the Heads of Executive Departments and Agencies, Subject: Open Government Directive," M 10-06, December 8, 2009, http://whitehouse.gov/open/documents/open-government-directive.

9. "Negotiate Health Care Reform in Public Sessions Televised on C-SPAN," Politifact.com, http://www.politifact.com/truth-o-meter/promises/obameter/promise/517/health-care-reform-public-sessions-C-SPAN/.

10. Alon Peled, "When Transparency and Collaboration Collide: The USA Open Data Program," *Journal of the American Society for Information Science and Technology* 62, no. 11 (November 2011): 2085.

11. Harlan Yu and David G. Robinson, "The New Ambiguity of 'Open Government,'" *UCLA Law Review* 59, no. 6 (August 2012): 178.

12. Jim Harper, "Grading Agencies' High-Value Data Sets," *Cato@Liberty* (blog) February 5, 2010, http://www.cato-at-liberty.org/grading-agencies-high-value-data-sets/.

13. Jim Harper, "Is Government Transparency Headed for a Detour?" *Cato@Liberty* (blog), January 15, 2010, http://www.cato-at-liberty.org/is-government-transparency-headed-for-a-detour/.

14. Yu and Robinson, p. 182.

15. Open Government Partnership, http://www.opengovpartnership.org/.

16. Office of the Press Secretary, The White House, "Fact Sheet: The Open Government Partnership," September 20, 2011, http://www.whitehouse.gov/the-press-office/2011/09/20/fact-sheet-open-government-partnership.

17. "John Boehner Introduces the House GOP Congressional Transparency Initiative," http://www.youtube.com/watch?v=hDr70qRv_9k.

18. Library of Congress, "About THOMAS," http://thomas.loc.gov/home/abt_thom.html.

19. The text of the House rules package is available at http://www.rules.house.gov/News/PRArticle.aspx?NewsID=33.

20. Committee on House Administration, Legislative Data and Transparency Conference, http://cha.house.gov/about/contact-us/legislative-data-conference.

21. Andrew Weber, "Introducing Congress gov!" *Custodia Legis: Law Librarians of Congress* (blog) September 19, 2012, http://blogs.loc.gov/law/2012/09/introducing-congress-gov/.

22. See Jim Harper, "Congress on Transparency: 'Needs Improvement,'" *Cato@Liberty* (blog), September 23, 2011, http://www.cato-at-liberty.org/congress-on-transparency-needs-improvement/; and Jim Harper, "Government Spending Transparency: 'Needs Improvement' Is Understatement," *Cato@Liberty* (blog), December 14, 2011, http://www.cato-at-liberty.org/government-spending-transparency-%E2%80%98needs-improvement%E2%80%99-is-understatement/.

23. Jim Harper, "Publication Practices for Transparent Government," Cato Institute Briefing Paper no. 121, September 23, 2011, http://www.cato.org/publications/briefing-paper/publication-practices-transparent-government.

24. See Jim Harper, "Congress on Transparency:

‘Needs Improvement’”; and Jim Harper, “Government Spending Transparency: ‘Needs Improvement’ Is Understatement.”

25. U.S. Senate, “Senators of the 112th Congress,” http://www.senate.gov/general/contact_information/senators_cfm.cfm.

26. U.S. Senate, “Committees,” http://www.senate.gov/pagelayout/committees/d_three_sections_with_teasers/committees_home.htm.

27. U.S. House of Representatives, “Document Naming Conventions,” http://cha.house.gov/sites/republicans.cha.house.gov/files/documents/committee_docs/CommitteeRepository-NamingConventions-v1-2-1.pdf.

28. U.S. Senate, “Daily Digest Committee Meetings/Hearings Schedule,” http://www.senate.gov/pagelayout/committees/b_three_sections_with_teasers/committee_hearings.htm.

29. U.S. Government Printing Office, “Congressional Reports,” http://www.gpo.gov/fdsys/browse/collection.action?collectionCode=CRPT.

30. U.S. House of Representatives, Office of the Clerk, “Legislation and Reports,” http://clerk.house.gov/legislative/legvotes.aspx.

31. U.S. Senate, “Recent Votes,” http://www.senate.gov/pagelayout/legislative/a_three_sections_with_teasers/votes.htm.

32. U.S. Office of Management and Budget, “Circular A-11,” http://www.whitehouse.gov/omb/circulars_a11_current_year_a11_toc.

Committee on Oversight and Government Reform
Witness Disclosure Requirement – "Truth in Testimony"
Required by House Rule XI, Clause 2(g)(5)

Name: Jim Harper

1. Please list any federal grants or contracts (including subgrants or subcontracts) you have received since October 1, 2010. Include the source and amount of each grant or contract.

none

2. Please list any entity you are testifying on behalf of and briefly describe your relationship with these entities.

The Cato Institute, Director of Information Policy Studies

3. Please list any federal grants or contracts (including subgrants or subcontracts) received since October 1, 2010, by the entity(ies) you listed above. Include the source and amount of each grant or contract.

none

I certify that the above information is true and correct.

Signature:

Date: March 12, 2013

Jim Harper
Director of Information Policy Studies

As director of information policy studies, Jim Harper works to adapt law and policy to the unique problems of the information age, in areas such as privacy, telecommunications, intellectual property, and security. Harper was a founding member of the Department of Homeland Security's Data Privacy and Integrity Advisory Committee and he recently co-edited the book *Terrorizing Ourselves: How U.S. Counterterrorism Policy Is Failing and How to Fix It*. He has been cited and quoted by numerous print, Internet, and television media outlets, and his scholarly articles have appeared in the Administrative Law Review, the Minnesota Law Review, and the Hastings Constitutional Law Quarterly. Harper wrote the book *Identity Crisis: How Identification Is Overused and Misunderstood*. Harper is the editor of Privacilla.org, a Web-based think tank devoted exclusively to privacy, and he maintains online federal spending resource WashingtonWatch.com. He holds a J.D. from UC Hastings College of Law.