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 tran-

sition 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.

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

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

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.

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

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

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 por-

table-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 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 per-

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

ceives 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 and , 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.

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.

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."

RECENT STUDIES IN THE BRIEFING PAPER SERIES

120. **Fannie, Freddie, and the Subprime Mortgage Market** by Mark Calabria (March 7, 2011)

119. **Short Sales Bans: Shooting the Messenger?** by Laurence Copeland (September 14, 2010)

118. **The Case for Auditing the Fed Is Obvious** by Arnold Kling (April 27, 2010)

117. **Scientific Misconduct: The Manipulation of Evidence for Political Advocacy in Health Care and Climate Policy** by George Avery (February 8, 2010)

116. **The Citizens' Guide to Transportation Reauthorization** by Randal O'Toole (December 10, 2009)

115. **ObamaCare: A Bad Deal for Young Adults** by Aaron Yelowitz (November 9, 2009)

114. **All the President's Mandates: Compulsory Health Insurance Is a Government Takeover** by Michael F. Cannon (September 23, 2009)

113. **High-Speed Rail Is Not "Interstate 2.0"** by Randal O'Toole (September 9, 2009)

112. **Massachusetts Miracle or Massachusetts Miserable: What the Failure of the "Massachusetts Model" Tells Us about Health Care Reform** by Michael Tanner (June 9, 2009)

111. **Does the Doctor Need a Boss?** by Arnold Kling and Michael F. Cannon (January 13, 2009)

110. **How Did We Get into This Financial Mess?** by Lawrence H. White (November 18, 2008)

109. **Greenspan's Monetary Policy in Retrospect: Discretion or Rules?** by David R. Henderson and Jeffrey Rogers Hummel (November 3, 2008)

108. **Does Barack Obama Support Socialized Medicine?** by Michael F. Cannon (October 7, 2008)

Published by the Cato Institute, Cato Briefing Papers is a regular series evaluating government policies and offering proposals for reform. Nothing in Cato Briefing Papers should be construed as necessarily reflecting the views of the Cato Institute or as an attempt to aid or hinder the passage of any bill before Congress.

CATO INSTITUTE

Contact the Cato Institute for reprint permission. Additional copies of Cato Briefing Papers are $2.00 each ($1.00 in bulk). To order, call toll free (800) 767-1241 or write to the Cato Institute, 1000 Massachusetts Avenue, N.W., Washington, D.C., 20001; phone (202) 842-0200; or fax (202) 842-3490. All policy studies can be viewed online at www.cato.org.