HARVARD | BUSINESS | SCHOOL

9-610-075
REV: MAY 13, 2010

KARIM R. LAKHANI
ROBERT D. AUSTIN
YUMI YI

Data.gov

The way to make government responsible is to hold it accountable. And the way to make government accountable is make it transparent so that the American people can know exactly what decisions are being made, how they're being made, and whether their interests are being well served.[1]

—Barack Obama, 44th President of the United States

Late in the afternoon one very wintry Thursday, in the middle of Washington D.C.'s famous "Snowpocalypse" snowstorm, in February, 2010, Vivek Kundra, Chief Information Officer (CIO) of the United States of America (USA), sat at his computer screen browsing datasets recently added to the Data.gov online website. Less than one full year into the project, more than 118,000 datasets had been published to the site. Despite this encouraging progress, Kundra had begun a new push aimed not just at greater numbers but also at datasets of especially high value to people *outside* government.

Like many big ideas, the idea behind Data.gov was simple: *Move government data to the web*—as much as could be released without compromising national security or individual privacy. *Make data available in raw, machine-readable format*—not finessed or massaged or spun, but ready to be processed, recombined, mashed up, and displayed visually. *Encourage people to use the data*—in any of the infinite number of inventive ways ingenious citizens might dream up, potentially unleashing new innovations and business ideas. *Harness the wisdom of crowds, couple that with transparency*—to do things government employees might not think to do, to achieve objectives far beyond those of government organizations. And, from the resulting potent mix—*get better government*.

Part of the broader "Open Government Initiative" launched by directive of the President of the United States on January 21, 2009, his first day in office, Data.gov had been on a fast track from the start. The first version of the website had gone live just a few weeks later, in May 2009. The initiative had received much public attention and press coverage, nearly all of it positive. Yet there remained an immense amount to do.

This was not the first government program aimed at making data available to the public, of course. Agencies such as the Bureau of Labor Statistics (BLS) had been providing great quantities of important data to people and organizations outside government for a very long time. But by a

[1] Barak Obama (January 21, 2009). Remarks by the President in Welcoming Senior Staff to the White House. *The White House*. Retrieved May 13, 2010 , from http://www.whitehouse.gov/the_press_office/RemarksofthePresidentinWelcomingSeniorStaffandCabinetSecretariestotheWhiteHouse.

Professor Karim R. Lakhani of Harvard Business School, Professor Robert D. Austin of Copenhagen Business School, and Columbia University Graduate Student Yumi Yi prepared this case.

considerable margin, Data.gov was more ambitious. It shifted the presumption about government data. If there was no good reason data had to stay secret, it should be published in machine-readable format so others could easily do whatever they wanted to with it. Veterans of data publication in government agencies, however, worried that raw data might be misunderstood or misinterpreted without vital context information. Indeed, everyone involved in the project could think of many legitimate reasons to hesitate in pursuit of this mission.

But Kundra intended to move fast. He kept close in mind key words in the President's directive: "…[create] an unprecedented level of openness in government…ensure the public trust and establish a system of transparency, public participation and collaboration…Openness [to] strengthen democracy, and promote efficiency and effectiveness in government" (**see Exhibit 1**). Often, he reminded his staff (and anyone else who would listen) that the data they meant to publish did not belong to the President, the Congress, or the Supreme Court. It belonged to the citizens of the United States, and they had an important duty to deliver that data to its rightful owners.

Alone in his office, on this Tuesday afternoon, however, listening to the sounds of snow removal equipment whirring and struggling to move mountains of white stuff outside just outside his window, he set aside noble objectives for a moment to focus in on practical challenges. He had his own mountains (of data) to move. Kundra knew the challenges. He'd done something like this before, first in Arlington County, Virginia, then as Chief Technology Officer of the District of Columbia (D.C.). Working for the energetic, reformist Mayor of D.C., Adrian M. Fenty, he served approximately 600,000 residents of D.C. and a larger population that swelled every weekday to nearly six million, as residents of surrounding areas converged on the city to help run the country. It was here that Kundra had pushed his first bold experiments in government transparency. And in doing so, he had discovered a practical truth: just about everyone approves of transparency in the abstract; but transparency in practice is almost always perceived as a threat by someone. In D.C., when Kundra published data that provided visibility into practices of government vendors, a controversy had erupted. "I thought that one would get me fired," Kundra recalled.

His program had worked in D.C. But would it, Kundra wondered, work at a national level?

Background

The view into federal government organizations has often been far from transparent. Practical difficulties in making available large quantities of government data play a major role in creating this problem. In the past, systems for distributing data were often assigned low priority and were thus nonexistent or unreliable. Many agencies held hundreds of thousands of data items but did not have information managers who could provide this data to the public in formats they could easily use (e.g., machine-readable). A citizen who wanted to access most government data needed prior understanding of how federal organizations and their data and systems were organized.

Historically, there had been little motivation to change this state of affairs. As organizations that carry out policies opposed by minority constituencies, who are thus prone to criticize execution of those policies, and that also have responsibilities to safeguard certain kinds of data (national defense secrets), federal departments and agencies have had natural inclinations toward secrecy. Classification restrictions, intended to safeguard data that needs to be secret, have tended to be applied too broadly, as government employees "played it safe" (when in doubt, classify it). The culture of government organizations has not emphasized or encouraged data sharing with the public.

The effects of this lack of transparency can be damaging, however. Citizens, whose tax dollars are being spent when government employees make decisions, have not been able to see *how* their money

is being spent, which has undermined confidence in government action. Even worse, poor visibility into government activity has limited the ways in which the public might engage in the work of their own government, contributing to and shaping policymaking and the execution of policies. Such engagement has been widely recognized as adding to the vibrancy of a healthy democracy.

To address these concerns, on September 26, 2006, President George W. Bush signed into law the Federal Funding Accountability and Transparency Act, also known as the Coburn-Obama bill, or, less formally "Google for government." It was an early effort of two new senators to bring transparency to government spending and provide public access to all government contracts.

Memorandum on Transparency and Open Government

On January 21, 2009, President Obama issued a Memorandum on Transparency and Open Government and called for recommendations for making the Federal government more transparent, participatory, and collaborative. The memorandum established three major objectives (**Exhibit 1**):

1. Government should be transparent.
2. Government should be participatory.
3. Government should be collaborative.

The Memorandum also led to the appointment, on March 5, of Vivek Kundra as Chief Information Officer of the U.S. Federal government. Along with Beth Noveck, later appointed Deputy Chief Technology Officer for Open Government, Kundra would lead the information transformation in the U.S. government.

The Open Government initiative that ensued following the memorandum would have huge implications in a number of ways. Ellen Miller, co-founder and executive director of the non-profit government accountability and transparency advocacy group Sunlight Foundation, highlighted changes in the way the US government would operate:

> Core to the President's campaign for government transparency is the use of technology in ways that redefines what 'public information' means – that is online information, information that is as easily searchable as it is easily accessible. This sea changing initiative aims to engage citizens with government in a more direct fashion. Thomas Jefferson is once supposed to have said that "information is power" and the Open Government initiative in this context empowers citizens by providing information online, in turn enabling citizens in their cities and towns and communities throughout the country to become powerfully informed.. Technology, of course is key to this, enabling mash ups of data relevant to their lives, offering economic possibilities heretofore not conceived of, allowing them to demand accountability of government and as a medium of distribution of the information itself. Through the Open Government Initiative, citizens would have access to information they care about that affect their lives while simultaneously participating in government every day -- not just on election day.

Vivek Kundra

Born in New Delhi, India, Kundra moved with his family to Tanzania when he was two years old, then to Washington D.C. when he was eleven, acquiring an eclectic mix of native languages, English, Hindi, and Swahili. He obtained undergraduate and masters degrees from the University of Maryland, and dabbled in a few technology based entrepreneurial ventures before beginning his career in public service in an unusual way on an infamous day:

> It was September 11, 2001 at about 8:30 in the morning and I had an interview for a position in Arlington County, Virginia, "Director of Infrastructure Technology"…in the middle of that interview, somebody knocked on the door and said "We've entered a Federal Emergency," and at that point we turned on CNN and saw the second plane go into the World Trade Center. Out the window, you could literally see the smoke coming from the Pentagon. I was interviewing with a panel, and we ended the interview, and somehow in the middle of all that my career in public sector began.

He spent two years working on infrastructure for Arlington County, just south of Washington D.C., at an intense time when people had become very worried about its security. He left that position to join a startup company, which was eventually sold, and then moved to another cyber-security related startup. His work with that company led to a meeting with then Lieutenant Governor Tim Kane, who would later become governor of the state of Virginia. Kundra and Kaine continued to interact, and in January 2006, Governor Kaine appointed Kundra to a dual role, as an Assistant Secretary of Commerce and Technology. In this position, he worked on broadband connectivity in rural areas, and also launched an initiative that foreshadowed things to come.

Governor Kaine placed high priority on creating opportunities for small businesses, especially those owned by historically disadvantaged groups such as women and minorities. To accomplish this, he directed that 40 percent of the state's discretionary expenditures (a component of the state's overall budget) should go to businesses owned by women or minorities. At first, government agencies self-reported their progress against this objective, but before long Kundra realized he could provide another view of this progress. Using a raw data file that contained raw credit card information and data from the state's procurement system, he created the "Small, Women and Minorities" (SWAM) dashboard (See **Exhibit 2** for 2010 data). The result was surprising, as Kundra explained:

> According to the report generated from raw data, we were under 4 percent as far as SWAM spending was concerned. Based on this, the Governor could call in cabinet secretaries and say 'Hey, what's going on here? The Executive Order says 40 percent.' Now, or last time I checked, we were at 43 percent. That's when I realized the power of transparency for the first time.

This story of government officials being called to account also, however, foreshadowed the potential controversy within what quickly became, for Kundra, a relentless focus on technology-enabled transparency.

When he left Virginia to join the administration of Washington D.C. Mayor Fenty in May 2007, it did not take him long to stir more controversy. A simple application that crossed crime data with GPS location data drew protests from police officials and some business owners.[2] Another application that showed government contracts awarded, the government official who awarded them, the hourly amount consultants were paid, and the period of the contracts generated even more furor, as it provided unexpected insights into the practices of vendors and the decisions of government officials. Not all applications stirred controversy; some helped visitors to D.C. design historic tours; others told residents where the nearest post office was wherever they were in the District. Kundra pushed over 400 real-time data feeds, and made the idea of delivering government data to the public from a single catalog seem normal (See **Exhibit 3**). Citizens in the District of Columbia used this information to track contract awards, crime incidents, vacant properties, construction projects, and many other things.

---

[2] See http://maps.google.com/?q=http://data.octo.dc.gov/Gateway_2008112000003.ashx?name=http://data.octo.dc.gov/feeds/crime_incidents/crime_incidents_current.kml.

Kundra also staged an "Apps for Democracy" competition. "I wanted to solve problems rapidly," he explained, "not in a two or three year cycle." To do this, he conjectured that there might be many creative people not working for the government who would be willing to help solve some of those problems. So he obtained a $50,000 budget and announced a contest that would award about half that amount in prizes to people who created the best applications based on the data his department had been making available. "I expected to get maybe 10 entries," Kundra said, "but we got 47 apps in 30 days." The $50,000 spent on the contest saved an estimated $2.6 million compared to what it would have cost to hire contract developers to do the same work.

In addition, he introduced a new model to manage technology investments in the District. His idea was to manage IT as a portfolio of stocks, with each project as a company, its team as the management, its schedule and financial status captured in market reports, and customer satisfaction as the market reaction. Applying stock-market analysis to government technology, Kundra could identify problems early in a project and take corrective actions, sometimes killing a project rather than allowing it to continue to waste resources.

As he experimented and gained experience with these approaches, he gained knowledge about how best to accomplish his objectives. He evolved a philosophy about sources of data: "We try to get data as close as possible to the source and in as atomic a form as possible, at the lowest possible level, without violating privacy or security, of course." He learned to move more slowly when ideas had potential to inspire controversy. He established a tendency to start small, trying a project with one agency before expanding it to others.

After the 2008 election, Kundra began working with the Obama transition team. Eventually, this work evolved into his next job, as federal CIO. Upon his departure, Mayor Fenty remarked: "Mr. Kundra has made the District the model for interactive citizen engagement through live data feeds and datasets."[3]

Open Source, User-Driven Innovation, Web 2.0, and the Wiki-Government Movement

"Sunlight is the best disinfectant"

— former US Supreme Court Justice Louis Brandeis

Kundra's experiments in transparency connected with a broader zeitgeist, democratizing attitudes and beliefs that had gained momentum in other domains of activity. Though it would be difficult to specify the beginnings and outlines of these movements, some of the important highpoints would include:

- Open Source Software – The emergence of the Linux computer operating system in the mid 1990s came as a revelation to many. No company led the development of this product, an extremely robust, high-quality computer system that outperformed many commercial products of its kind; rather, individual volunteers and employees from many organizations donated their time and efforts in a way, that, in the words of open source guru Eric Raymond, "coalesce[d] as if by magic out of part-time hacking by several thousand developers scattered all over the planet, connected only by the tenuous strands of the Internet."[4] To many, this remarkable achievement was the first indication that the Internet had changed something

[3] District of Columbia Mayor's Office (November 12, 2008). Maryland Tech Council Honors CTO Vivek Kundra for Outstanding Technology Leadership. *DC.gov*. Retrieved May 13, 2010, from http://newsroom.dc.gov/show.aspx/agency/octo/section/2/release/15435.

[4] See *The Cathedral and the Bazaar*, by Eric Raymond, http://catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/.

dramatic about the way things could be produced—that it somehow permitted communities of people with varied expertise to produce better outcomes than a single firm working on a proprietary offering.

- Wikipedia and the Wisdom of Crowds – In the early 2000s, Jimmy Wales and Larry Sanger created an on-line encyclopedia that anyone could add to or change. Remarkably, this completely open encyclopedia evolved into a comprehensive resource with accuracy that rivaled the best of the traditional encyclopedias. The key to this accuracy was a principle first identified by Eric Raymond, which he stated in a geeky but memorable way: "Given enough eyeballs, all bugs are shallow."[5] By this he meant that crowds (with their many, many eyeballs) would invariably spot and correct errors when they were present in anything that came to the attention of the crowd. The general principle, that crowds were often wiser than individuals when it came to many tasks, could be applied to many domains of activity. Various business models emerged at companies like InnoCentive and TopCoder, wherein crowds competed in contests to win cash prizes by solving tough technical problems that companies could not solve for themselves internally.

- User Innovation – Since the mid 1980s, Eric Von Hippel, a professor at MIT had noted that users often generated functionally novel innovations at the point of use of a technology. Von Hippel argued that manufacturers should treat these user innovations as a genuine source of product and service innovation. Since that time, von Hippel and his academic colleagues have shown that this is an extremely viable approach in many contexts, and that the Internet has enlarged the opportunities for "democratizing innovation" (as von Hippel calls this approach in his own book by the same name).[6]

- Social media, Web 2.0, search technologies, smartphones – All of these movements benefited from and interacted with new technologies that emerged in the mid-2000s. Social media (sometimes called "Web 2.0 technologies), such as Facebook, blogs, and Twitter, became means of organizing communities. Search technologies (Google leading the way) made it possible to find things that had been hidden away and difficult to access, thus changing the way people managed knowledge (why remember it if you can always search for it?). Smartphones added a final piece of the puzzle, by providing people with a device they could keep with them always, that linked them to online communities, and gave them access anytime, anywhere to the benefits of being online, in essence creating a mobile internet.

In April 2009, Beth Simone Noveck published *Wiki Government: How Technology Can Make Government Better, Democracy Stronger, and Citizens More Powerful*, which argued that these ideas that were already at work democratizing many domains of human activity should be harnessed to the goals of improving democracy itself. In September 2009, at the Gov 2.0 conference in Washington, D.C., John Podesta, former chief of staff to President Bill Clinton, argued that emerging communications technology should be a "tool of empowerment" for citizens in societies where political expression is constrained. Understanding the challenges and opportunities for governments, policy makers outside government, and those traditionally involved in providing oversight, was, however, a significant task that had only just begun.

---

[5] Ibid.

[6] See *Democratizing Innovation*, by Eric Von Hippel, http://web.mit.edu/evhippel/www/democ1.htm.

The Data.gov Initiative

Kundra wasted no time in implementing his ideas about e-governance at the federal level. The Data.gov initiative was established by the Federal Chief Information Officers' Council and the E-Government and Information Technology Office at the Office of Management and Budget on March 11, 2009, with a memorandum that asked CIOs of government agencies to nominate datasets that were suitable for the Data.gov initiative. This initial call yielded 76 data sets and tools from 11 agencies. The idea was to build a single web address from which the public could quickly and efficiently access data of interest. Just having a single URL for this purpose would be a big break from the past. At the beginning of 2009, the US government had 24,000 domains with the .gov ending (GetOutDoorsItsYours.gov, for example, maintained by the Parks Service), all of which connected the public with 30 million web pages, all of which contained some kind of government information.

Building Data.gov

Kundra began Data.gov by asking for volunteers from the agencies, and he zeroed in on a crucial question: What datasets should we go after first? He knew there was potential to spend two years or more just debating privacy, security, and legal issues, but he wanted to avoid that. This meant going first after data that was not controversial but that would provide value to consumers of the information. Geospatial data from the Department of Interior (DOI) emerged as an early strong candidate. Some data from the Census Bureau, the Center for Disease Control (CDC), and the Environmental Protection Agency (EPA) also fit the profile. Kundra asked Sanjeev "Sonny" Bhagowalia, the CIO of the DOI, and Linda Travers, CIO at EPA, to co-lead the effort with other CIOs to nominate datasets that fit the profile and could provide value.

One of the first aspects that came under debate among the agencies was whether the data provided should be "wholesale" (raw data) or "retail" (displayed, explained, interpreted). Based on his past experiences, Kundra thought the wholesale, raw data approach was extremely important. Some agencies that had a lot of experience in providing information to the public worried, however, that raw data might be misinterpreted, and that this might result in adverse consequences. These were legitimate concerns. Rick Kryger, of the Bureau of Labor Statistics explained:

> You click on a link and all of a sudden you are getting hit with 100 or 200 megs via FTP download. Without people understanding some of the context of that data, there is potential for problems that might undermine confidence in our data…Statistical agencies have made pledges of confidentiality. That people are not going to have access prior to release time. The Consumer Price Index or the unemployment situation…by law, we must protect this data…Back in the 70s, President Nixon was not happy with some of the employment information that BLS was producing, it did not show the right trend, [so] he replaced BLS executives because he didn't like the way the numbers were coming out, or what we were saying in analysis of the numbers. So there are good reasons for these laws and policies. Some of this information can move markets, and getting it early would give someone unfair trading advantages.

Although Kundra argued the case for the wholesale model, he and the others in project recognized the need to be sensitive to issues of quality of information, especially where people out in the world already relied upon the data.

Another issue that came up early concerned the actual means in which Data.gov would be implemented. Early thinking aimed to establish a large data warehouse full of government data. This, however, generated many concerns about maintainability, how the data would be kept current and its quality assured. Agencies that produced high value data reports took extensive steps to assure the

currency and accuracy of their data, and they were reluctant to turn that data over to another entity without evidence of similarly extensive precautions. Ultimately such concerns were addressed when it was agreed that Data.gov would operate as an *index*, not as an actual warehouse for data. People seeking data might discover a dataset on Data.gov, but if they wanted to download it they would be redirected to the agency source of the data. Kundra pushed for agreement that even with this indexing, data would never be more than three clicks away from the top level catalog.

Another challenge concerned so-called "metadata"— information about the datasets such as descriptions, standardized "tags" that indicated what could be found in which fields, and in what formats. For the Data.gov platform to interoperate with itself and outside platforms, metadata needed to be standardized to the extent possible. Thus agencies were not only asked for datasets to link to, they were also asked for metadata about each dataset in a standard template. Different agencies came forth with datasets in different formats: Excel spreadsheets, Oracle database files, and written reports. The different data formats generated challenges for how the metadata template was designed, as it needed to accommodate all these formats. The metadata format would serve as basis for search, navigation through data, and combination of datasets, so it was of vital importance.

The Data.gov team used an "agile development" approach that would iterate on 24 to 48 hour cycles. They worked issues one at a time, as they appeared. They appeared daily, as Kundra described: "You might have statutory legal requirements that prevent data from being released. Or you might have a policy within in an agency that prevents information from being released. Or you might have a system on the backend not capable of handling the processing requirements." The IT work was done by vendors, under existing blanket contracts, for small amounts of money, which avoided a lengthy bid process. The development concept, as Sonny Bhagowalia put it, was "start small, think fast, get something on, build it, incrementally, see how the thing is. Figure out a way to get it out."

Throughout the process, they employed great care to avoid violations of privacy or harm to national security. Government data was organized in *enclaves*: Unclassified, Sensitive but Unclassified, Secret, Top Secret, Top Secret SCI. Data.gov dealt only in unclassified information, but it actually posed stiffer restrictions than that. Information to be included in Data.gov had to be unclassified, *and* contain no person identifying information, *and* contain no national security information *and* meet the guidelines of the Information Quality Act (IQ guidelines).

Data.gov was launched on May 21, 2009, on the 120-day anniversary of the presidents Open Government Memorandum to serve as the single access point for publicly available authoritative Federal data (See **Exhibit 4** for website).

Transparency, Participation, Collaboration

As launched, Data.gov's vision echoed the president's Memorandum on Open Government with its focus on transparency, participation, and collaboration (See **Exhibit 5** for *Guiding Principle)*.

Transparency

At the core of Data.gov was the intent to make Federal sector data more accessible and usable. Increasing the ability of the public to discover, understand, and use the vast stores of government data would increase government accountability and unlock additional economic and social value. Data.gov would offer coordinated and cohesive cross-agency access to data and tools via a non-agency specific delivery channel. In doing this, it would enhance the ability of developers, researchers, businesses, and the general public to find information by offering metadata catalogs integrated across agencies, and it would eventually provide the opportunity for agencies to leverage

Data.gov shared data storage services if they so desire. A more consolidated source for data and tool discovery would allow the public to navigate the Federal sector data holdings without having to know, in advance, how Federal agencies and data programs are organized. These capabilities would increase transparency by enhancing the discoverability of specific data and information from the tools made available through Data.gov.

Participation

Public participation was a key pillar of the open government agenda and was critical to the success of Data.gov. The site provided mechanisms for the public to participate in the evolution of the site. Future versions of the site would allow the public to post, rate, and prioritize suggestions. Data.gov increased the opportunity for the public to discover and understand the data resources available and, for those with the inclination, to subsequently build applications, conduct analyses, and perform research. A basic value proposition of Data.gov was to spur additional analysis and innovation. Data.gov sought to engage the public in expanding the creative use of Federal data beyond the walls of government by encouraging the development of innovative ideas (e.g., web applications); combining Federal and other data to gain new insight into efficiency and effectiveness of government; pursuing new economic and socially-based ventures (new businesses, for example); thereby enriching the lives of citizens.

Collaboration

Key collaboration mechanisms to improve and evolve Data.gov included direct feedback, comments, and recommendations from the public. For example, individuals were encouraged to suggest datasets to add to Data.gov, rate and comment on the value and quality of current datasets, and suggest ways to improve the site overall. The project also aimed to make use of crowd-sourcing techniques. These could enable community-based creation of new datasets or the ability to tag local landmarks or points of interest in geospatial data, for example. Beyond these Data.gov-centric collaboration concepts, Data.gov could be a source of data for non-government sites that developed their own forums for collaboration.

Management Systems and Processes

Data.gov operated at two levels. The website was the public presence, delivering on the government's commitment to transparency. On the policy level, Data.gov was about increasing access to data that agencies already made available and making available additional data sources that had not been freely presented to the public in the past.

For data that were already available, the emphasis was on improved search and discovery, as well as provisioning of data in more usable formats. For data that had not been widely available, the focus was on providing data in a more timely and granular manner while still protecting privacy, confidentiality, and security.

On an operational level, Data.gov's focus was creating the website and associated architecture designed to catalog Federal datasets, improve search capabilities, and publish information designed to allow the end user to determine the fitness for use of a given dataset for a particular application. The goal was to create an environment that fostered accountability and innovation. Realizing the vision for the website required agencies to:

- Make their most relevant and informative data and related presentation tools available through Data.gov

- Do so in a manner that supports use and innovation by stakeholders – public or private
- Agree on a shared performance management framework centering on quantifying the value of dissemination of high quality, secure, public information that does not raise privacy or confidentiality concerns.

As the numbers of datasets grew, the process for moving datasets onto Data.gov needed to become increasingly formal and efficient. This lead to the development of the "Dataset Management System" (DMS) which, in the words of its project manager, Marion Royal:

> Provides a complete line of sight into a dataset being moved to the web. The originator of the dataset, metadata, the agency point of contact, etc. We validate the dataset, it goes to QC, and eventually we publish it…If I look in DMS, I can see what data is in draft, what's under review, what is at the Program Management Office, what is in the final queue, and what is published. And I can provide visibility into all that for any manager that needs to see the whole view.

Along with the new process came new governance procedures. The Data Architecture Subcommittee assisted with design of the metadata template. The Architecture and Infrastructure Committee helped with the physical arrangements. The CIO Council, composed of all agency CIOs, had oversight over the total process.

Data.gov Applications in Action

Recovery.gov was the U.S. government's website for reporting to the public on use of the $787 billion authorized for use by the Recovery Act, a response to the difficulties in the banking sector that began in September 2008. Kundra worked on the Recovery.gov site upon arriving in his new role to make sure that it operated in an intuitive manner.

Another site, the IT Dashboard, tracked $76 billion in IT investments made by the Federal government. The site showed each investment, and an abundance of information about it such as who won the contract and where the project is in terms of cost and schedule compared to plan. The site both displayed this information and allowed raw data underlying displays to be easily downloaded. Kundra used the IT Dashboard for the same kind of project oversight approach he'd used in D.C.:

> One project was 17 months behind schedule and 110% over budget…We halted it because there wasn't enough planning, there wasn't enough infrastructure…When we talk about killing projects, there is usually a need, a good reason the project was initiated. But one of the reasons we went live with the IT Dashboard was so that everybody could see where we are, so we could create Darwinian pressure…If an investment isn't paying off, we divest, freeze, get a new team in play.

A program developed in 90 days with the US Citizenship and Immigration Service allowed green card applicants to go online and sign up to get text message updates on the status of their applications. Applicants could track their materials as they moved through each stage of the process. The system also allowed applicants to query processors during processing, and showed average processing times in each office, Austin versus Houston, or New York versus Chicago.

Another initiative created benefits for taxpayers by sharing information between the Internal Revenue Service (IRS) and the Department of Education. Every year, most people who have a student in college in the US need to fill out the Department of Educations Free Application for Federal

Student Aid (FAFSA) to qualify for financial assistance in paying for their university education. Virtually all colleges and universities in the US require the form, which standardizes reporting to colleges on a student's family's income, allowing everyone to compete for available need-based aid on an equal basis. Almost all of the information required on the FAFSA form is already on file with the IRS, as part of a family's annual tax filing. In fact, the Department of Education requires that colleges and universities verify information on the FAFSA form, which they do by providing a copy of their annual tax returns to the universities. All this could be vastly simplified if the IRS provided the necessary information directly to its sister agency in the federal government, the Department of Education. Historically, however, this sharing was a non-starter because of the IRS's strong policies to ensure the privacy and protection of taxpayer information. Working as part of the Data.gov program, however, Kundra and his staff made a breakthrough. As of January 2010, IRS data would automatically populate the FAFSA online form upon the taxpayers' authorization. Needless to say, this vastly simplified the FAFSA process, reduced error rates, and relieved colleges and universities of unnecessary data verification burdens.

Non-government Applications that Use Data.gov as a Source

An important value proposition of Data.gov is that it allows members of the public to leverage federal data for robust discovery of information, knowledge, and innovation. Making Federal data more transparent has many benefits inside the government, including the potential to maximize the return on investments in collecting and managing data in a manner that transcends agency stovepipes. But, as was the case for the public release of Global Positioning System (GPS) data by the US Department of Defense, releasing datasets beyond the walls of government also facilitates entrepreneurship and value creation that far exceed what government can do on its own.

The Sunlight Foundation organized the $25,000 Apps for America 2: The Data.gov Challenge in April 2009. "By setting government data free on its new Data.gov site, the Obama administration enabled and encouraged the creation of fresh, new ideas that could help citizens get more involved in their government," said Clay Johnson, director of Sunlight Labs. "Seizing upon this important moment, Sunlight organized this Apps for America contest to catalyze the development of useful applications and visualizations to make this information more comprehensible to more people. We also wanted to demonstrate to the government that when it makes its data available, it makes itself more accountable and creates more trust and opportunity in its actions." [7] The contest, in the space of three months, spurred the creation of 47 applications and the top three winners included:

- Datamasher.org: a web application that allows anyone – without expertise or knowledge of web programming – to choose different government data sets and mash them up to create visualizations on a state by state basis.

- Govpulse.org: a website that enables quick and easy search of the Federal Register, the official repository of all government agency and executive branch publications and notices, including location-based visualization.

- Quakespotter.org: a cross-platform desktop application that shows where earthquakes are happening and matches them to Twitter messages that mention the event.

The *Forbes* list of "America's safest cities" shows just the tip of an iceberg of possibilities. By drawing upon multiple government and external sources, it created something more valuable than

[7] The Sunlight Foundation (September 9, 2009). The Sunlight Foundation Names Apps for America2 Winners. *The Sunlight Foundation*. Retrieved May 6, 2010, http://sunlightfoundation.com/presscenter/releases/2009/09/09/sunlight-names-apps-america2-winners/

what could be produced by one source alone. To arrive at their determination of the safest cities, Forbes took information from the Bureau of Labor Statistics, the National Highway Traffic Safety Administration, SustainLane.com, the National Oceanic and Atmospheric Administration, the United States Geological Survey, the Department of Homeland Security, the Federal Emergency Management Agency, Risk Management Solutions, the Federal Bureau of Investigation, and Sperling's Best Places.[8]

This example, though powerful, is only the beginning of what is possible. Thought is being given to how Data.gov might evolve to become compatible with the semantic web or data web. This future would involve the adoption of protocols, and their implementation by agencies, to encode meaning into data in such a way that they are directly interpretable by computers. Instead of having "data on the web", there will be a "web of interoperable data". Through the data web, data aggregation and analysis might be done directly through machine interaction, and new applications and services might be more efficiently created.

A Convergence of Interests

One indicator of the importance of the Data.gov initiative was the interest in it shown by giants of the tech world. Companies like Amazon, Google, and Microsoft had indicated strong support for the project. Some of this support had to do with desires by company executives to contribute to the common good. But some of it related to future possibilities for profit making business. Ray Ozzie, Microsoft's Chief Software Architect, explained:

> It's in the best interests of our shareholders for us to have both platforms and apps that our customers are interested in. We think we can catalyze network effects by taking data made available by both Data.gov and commercial vendors and bringing it together into a platform that sets key constraints around it such as the normalization format and the programmatic access methods. Essentially, we take the burden of hosting, curating, serving, and reformatting some of the data. If we do a good job, the marketplace will find [our tools] more attractive than others. We're trying to create an ecosystem around these various types of data.

Microsoft has a project codenamed "Dallas"[9] aimed at executing just this sort of initiative. Standardizing access, data formats, and the like did not seem very glamorous in itself. But, as Ozzie further explained, getting this right could be a very big deal from a business standpoint:

> One of the key things that people don't quite understand is that computation and storage collocation makes a huge difference. If you have a government dataset that's sitting on government servers, and a commercial dataset sitting on the other side of the country or the other side of the world on a different set of servers, it's hard to join them to have one virtual view. But if they're both in the same data center and they've been synchronized or duplicated into the cache, and the app is also in that same data center, you can efficiently join all three in ways that you could not do otherwise.
>
> For example, imagine that the government published a dataset in which one of the columns is something related to geo-location, and that a commercial provider has another piece of data based on geo-location. Now let's say there is a third provider with an app that can display geo-location data as icons on a surface, like Google Maps or Bing Maps. If you've got

[8] Zack O'Malley Greenburg (October 26, 2009). America's Safest City. *Forbes.* Retrieved May 6. 2010, http://www.forbes.com/2009/10/26/safest-cities-ten-lifestyle-real-estate-metros-msa.html.

[9] See http://pinpoint.microsoft.com/en-US/Dallas.

everything collocated and in a common format, then you can get out more capabilities than you put in.

The vision these and other companies competed to realize involved joining, crossing, recombining or mashing up data from different sources in real time, then delivering the information, also in real time, to mobile devices.

Tim O'Reilly, founder of O'Reilly Media, noted that the innovation potential of marrying government data with business ingenuity was quite broad and reached beyond the public transparency goals. He gave the example of Crimereports.com, which provides a free public service listing of postal-code based crime reports to citizens but also offers tools to police departments that allow them to create back office reports tailored to individual police beats. O'Reilly observed:

> There are two sides to transparency and access to government data. There's public benefit to the data access. But frankly, all of the transparency groups, cool as the stuff they do is, are appealing to a tiny fraction of the public. A bigger opportunity lies in finding real opportunities and real markets for this data. We need to use it to improve the actual functioning of government and of markets. Creating tools that solve real problems for paying customers using this data is where a lot of the necessary innovation will come from. But it is still very early, and I would say that we are at a stage where if this were the Web, people have just discovered the "Blink" html tag, and a lot of things that are being done are not well thought out. There is, however, lots of entrepreneurial energy, and people are figuring out how to make markets with government data and how to make applications that pay for themselves. That is what is so exciting about the potential of Data.gov.

Although no one could yet see all the possibilities (or risks, for that matter), the opportunities around creating value-added products and services on top of government data was becoming apparent to both large and small firms. And large sources of rich and varied data would be key ingredients in the futures executives in these companies imagined. In the not too distant future, companies would build formidable business models from linkages to Data.gov and other data sources like it.

Challenges

Though excellent progress had been made in a short time, challenges remained. Agencies were very busy, and Data.gov had to be sensitive to creating additional burden that might interfere with other important work.

Privacy and national security issues were an ever-present concern. Although the Data.gov team was careful to avoid releasing information that might harm national security, there was always the possibility that seemingly harmless data might be combined with other seemingly harmless data to create harmful data. One example: A security researcher in the UK took seemingly harmless data from two different public sites (not government sites), linked this with other information about electric power generation, and published a paper that suggested ways of harming the US electric utility grid. Ellen Miller states:

> "It is important to build into the process a way to review data to insure that privacy is not violated or that other sensitive information is not released. There's a balance here – but the interests of a healthy democracy require nothing less than real time online access to government information. Of course, we do not want to release data that could cause any harm

but if we are prudent we can and should move forward aggressively because of the transformation opportunity that the Open Government Directive suggests."

You could interpret this story in different ways, however. The most straightforward interpretation was that data should not have been publicly available— an argument for more secrecy and more conservative judgments about what to publish. But another interpretation took a page from Eric Raymond's open source philosophy, "given enough eyeballs, all bugs are shallow." This suggested that the public information allowed us, using the large number of public eyeballs, to discover a threat that we could now fix. Without making the dataset available, only bad guys motivated to acquire this information for insidious purposes would have discovered the vulnerability in the power grid.

Microsoft's Ray Ozzie pointed to why he thought Data.gov needed to continue to focus on such concerns, but also why he thought such issues could be managed:

> Correlation is an amazing thing. I'm sure somebody argued that all the flight data out there right now would be a national security threat. By knowing where every plane is, for example, you might be able to track corporate executives. But by pulling back a little, by letting people opt out from providing information, maybe we can find the right balance. Having directory assistance but letting people make their number unlisted was an explicit decision.

Other worries also occurred to Ozzie, especially those related to data quality, partly because these also applied to what he and others were doing at Microsoft and elsewhere:

> Once people start using data, some very hard issues that many people don't yet understand will arise related to the provenance and quality of data that's released. The legal staff and the people who are doing the ingestion, cleaning and rendering of data will need to take on the additional burden of keeping that data synchronized and up to date. There are many different nuanced issues once you have production systems based on this.
>
> Transparency is always a double-edged sword…What if an agency publishes data, and we look at it and go "Wow, that's 1950s stuff" or "There are all these data entry errors."

Dave Campbell, a Microsoft Technical Fellow, added his own thoughts on this:

> How do you reason based on data when you take four sources of different levels of quality and mash them up and produce yet a different stream? And someone consuming that stream, how do they reason about the quality of what they're getting? Once someone is consuming it, what is the sort of guarantee or implied guarantee relative to its freshness? Who's going to be maintaining it five years hence when someone's got a valuable solution based on it?

O'Reilly raised the issue that the Data.gov initiative in many ways had the potential to be a significant platform for innovation, but in its current state it was more of a data index:

> I think the biggest flaw with Data.gov is that they have not built any sort of community formalisms that make developer platforms a huge success. If you think about when Amazon rolled out Amazon Web Services – they quietly found people who were working with Amazon data, and they brought them in and they showed them the cool new features they would be supporting, so there would be some working code at the launch. Or you look at the Apple iPad launch. There are developers who are in there working to build cool apps so that when Apple does the rollout, there's already critical mass there. The government needs to go beyond a data repository mentality to creating tools and services that enable others to innovate.

The Data.gov project, as might be expected, had also attracted some critics. OpenTheGovernment.org, a non-profit proponent of the Open Government initiative, expressed frustration with what they called "sluggish process." A spokesperson suggested that Data.gov should place emphasis on making internal records, such as policy papers and emails, available: "Data is important for accountability, but so is how policy was formed."[10] Alex MacGillis, a writer for the *Washington Post*, said of Recovery.gov, that it "offers little beyond news releases, general breakdowns of spending, and acronym-laden spreadsheets and timelines."[11]

Just a Beginning?

Kundra knew that there were literally millions of datasets in the possession of the US government. He had recently worked on early prototypes of the US Center for Medicare and Medicaid Services (CMS) dashboard that would allow for cost comparison across the nation for various health care products and services. Surely, he thought, such a tool on Data.gov would bring insight into the even more serious political storm over health care reform brewing in the USA. So they were really just beginning. But already, the launch of Data.gov had catalyzed similar initiatives across the United States and, indeed, internationally. Amid it all, Kundra tried to keep in mind the grand motivations behind the day-to-day work. He asked himself and his staff questions about their important mission:

> What do we do to introduce fundamentally game changing approaches in the 21st century to how the federal government works, and how do we do it in a transparent, open way? How do we ensure that the debates we're having in our political system are grounded in evidence and in data, and not in feelings and rhetoric? How do we make sure, as we think about a vibrant, healthy democracy, that it doesn't become corrupt, that it doesn't get owned by special interests, that it really serves the American people in the way that our founding fathers envisioned. Data.gov makes sure we don't get a government that's closed, secret, and opaque, that's become corrupt, that rots from the inside because only a few people have access to good information and they use it for their own self-interested purposes. And then, of course, there's innovation. We can drive innovation. We can spur innovation in ways we can't even imagine today. Not just in government, but also throughout business and society as a whole.

---

[10] Jesse Lee (May 21. 2009). Transparency and Open Government. *The White House.* Retrieved May 6, 2010, from http://www.whitehouse.gov/blog/09/05/21/Opening.

[11] Alec MacGillis (May 21, 2009). Tracking Stimulus Spending May Not Be As Easy As Promised. *The Washington Post.* Retrieved May 6, 2010, from http://www.washingtonpost.com/wp-dyn/content/article/2009/05/20/AR2009052003535.html.

Exhibit 1 Memorandum For the Heads of Executive Departments and Agencies

SUBJECT: Transparency and Open Government

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.

Government should be transparent. Transparency promotes accountability and provides information for citizens about what their Government is doing. Information maintained by the Federal Government is a national asset. My Administration will take appropriate action, consistent with law and policy, to disclose information rapidly in forms that the public can readily find and use. Executive departments and agencies should harness new technologies to put information about their operations and decisions online and readily available to the public. Executive departments and agencies should also solicit public feedback to identify information of greatest use to the public.

Government should be participatory. Public engagement enhances the Government's effectiveness and improves the quality of its decisions. Knowledge is widely dispersed in society, and public officials benefit from having access to that dispersed knowledge. Executive departments and agencies should offer Americans increased opportunities to participate in policymaking and to provide their Government with the benefits of their collective expertise and information. Executive departments and agencies should also solicit public input on how we can increase and improve opportunities for public participation in Government.

Government should be collaborative. Collaboration actively engages Americans in the work of their Government. Executive departments and agencies should use innovative tools, methods, and systems to cooperate among themselves, across all levels of Government, and with nonprofit organizations, businesses, and individuals in the private sector. Executive departments and agencies should solicit public feedback to assess and improve their level of collaboration and to identify new opportunities for cooperation.

I direct the Chief Technology Officer, in coordination with the Director of the Office of Management and Budget (OMB) and the Administrator of General Services, to coordinate the development by appropriate executive departments and agencies, within 120 days, of recommendations for an Open Government Directive, to be issued by the Director of OMB, that instructs executive departments and agencies to take specific actions implementing the principles set forth in this memorandum. The independent agencies should comply with the Open Government Directive.

This memorandum is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity by a party against the United States, its departments, agencies, or entities, its officers, employees, or agents, or any other person.

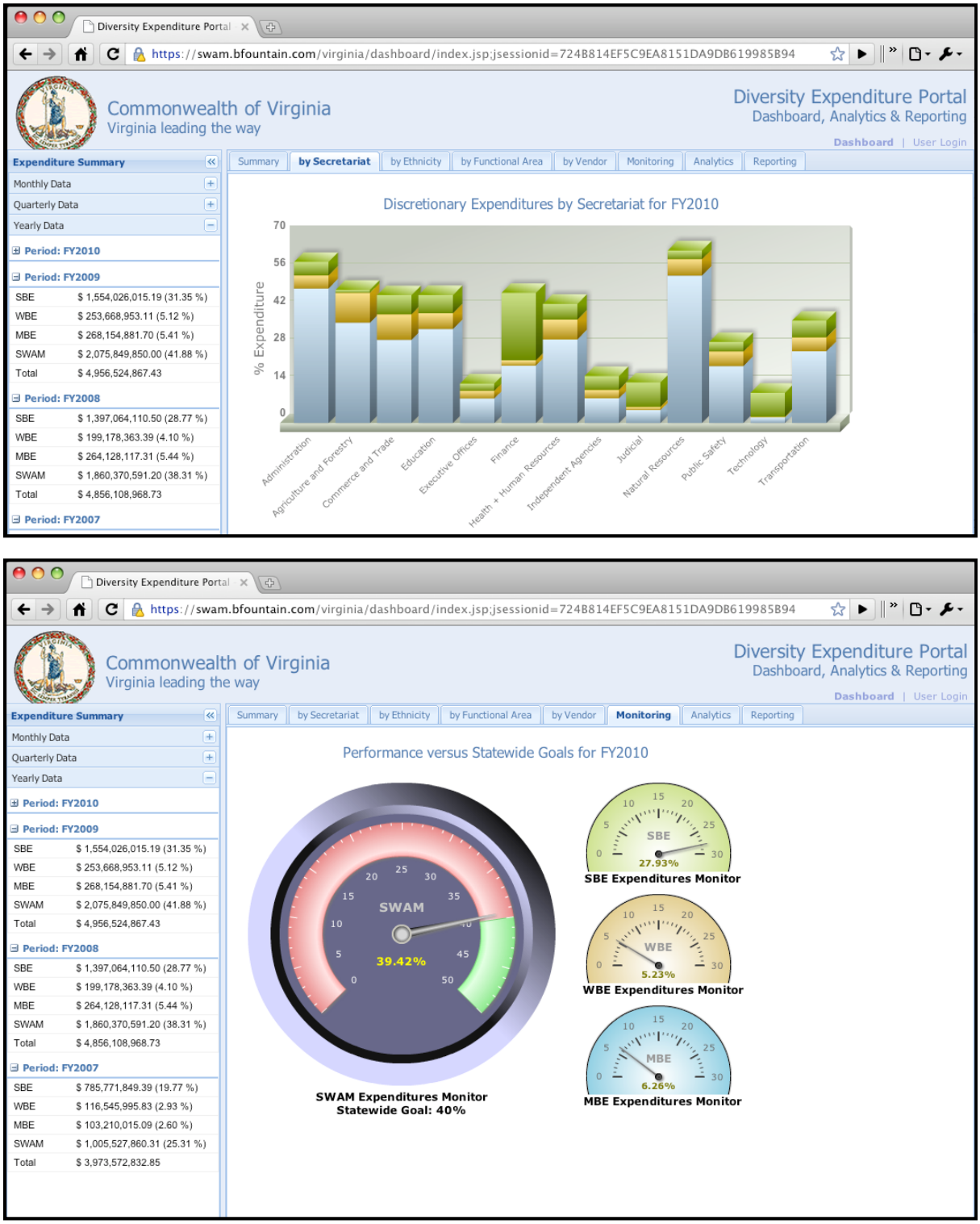
This memorandum shall be published in the Federal Register.

BARACK OBAMA

(Issued January 21, 2009)

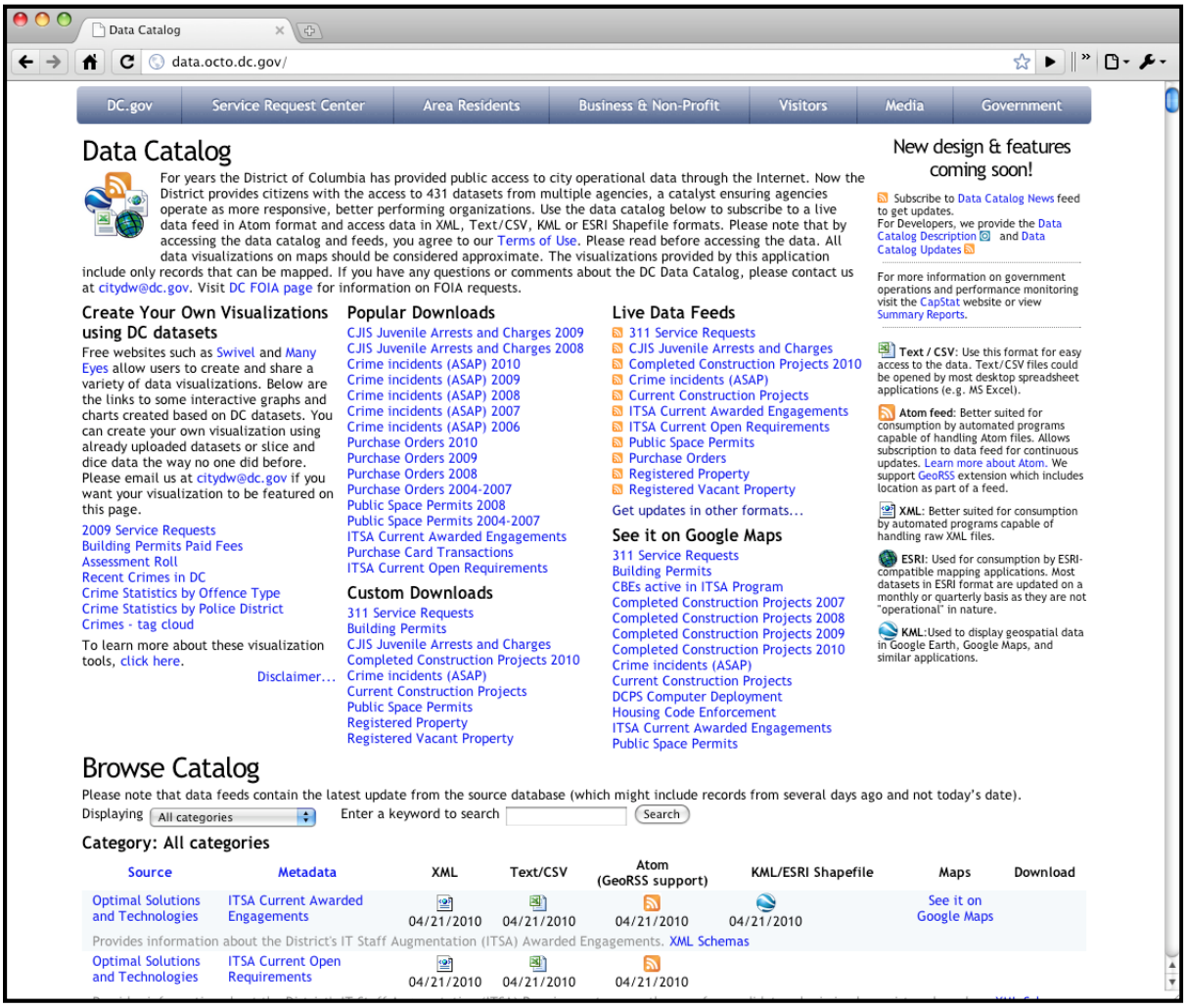
Source: http://www.whitehouse.gov/the_press_office/Transparency_and_Open_Government/, accessed May 2010.

Exhibit 2 State of Virginia SWAM Dashboard



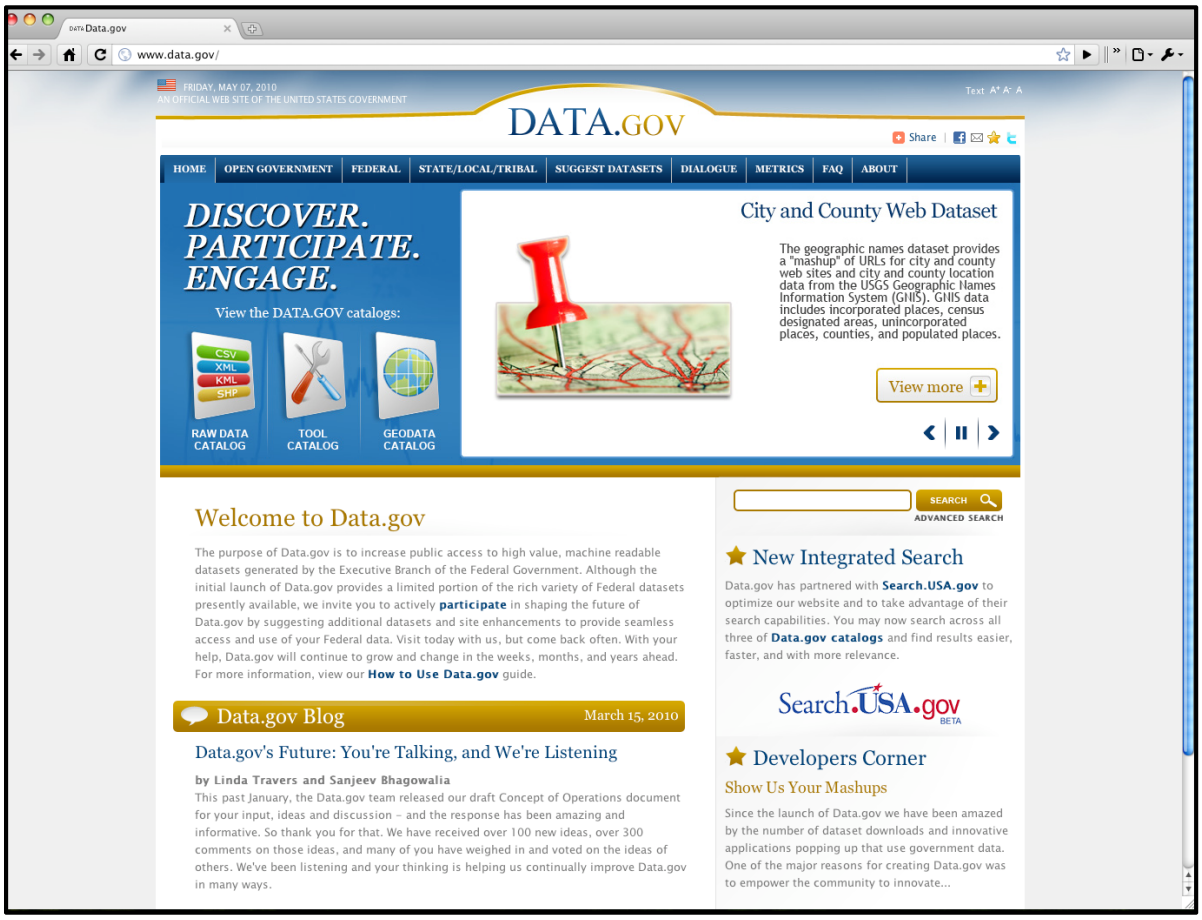
Source: https://swam.bfountain.com/virginia/dashboard/index.jsp, accessed April 2010.

Exhibit 3 Washington D.C. Data Catalog



Source: http://data.octo.dc.gov/, accessed April 2010.

Exhibit 4 Data.gov Website



Source: Data.gov, accessed May 2010.

Exhibit 5 Data.gov Guiding Principles

Key Data.gov guiding principles included:

1. Focus on Access

Data.gov was designed to increase access to government data as close to the authoritative source as possible. The goal was to strengthen democratic institutions through a transparent, collaborative and participatory platform while fostering development of innovative applications (e.g. visualizations, mash-ups) and analysis by third parties. Policy analysts, researchers, application developers, non-profit organizations, entrepreneurs and the general public should have numerous resources for accessing, understanding and using the vast array of government datasets.

2. Open Platform

Data.gov would use a modular architecture with application programming interfaces (API) to facilitate shared services for agencies and enable the development of third party tools. The architecture, APIs, and services would evolve based on public and agency input.

3. Disaggregation of Data

Data should be disaggregated from agency reports, tools or visualizations to enable direct access to the underlying data.

4. Grow and Improve through User Feedback

Feedback should be used to identify and characterize high value data sets, set priorities for integration of new and existing data sets and agency provided applications, and drive priorities and plans to improve the usability of disseminated data and applications.

5. Program Responsibility

Agency program executives and data stewards were responsible for ensuring information quality, providing context and meaning for data, protecting privacy, and assuring information security. Agencies were also responsible for establishing effective data and information management, dissemination, and sharing policies, processes and activities consistent with Federal polices and guidelines.

6. Rapid Integration

Agencies should rapidly integrate current and new data into Data.gov with sufficient documentation to allow the public to determine fitness for use in the targeted context.

7. Embrace, Scale and Drive Best Practices

Data.gov would implement, enhance and propagate best practices for data and information management, sharing and dissemination across agencies, with our state, local and tribal partners as well as internationally.

Source: CIO.gov, accessed May 2010.