Data.gov Concept of Operations

Office of E-Government and IT

Office of Management and Budget

Powered by the Federal Chief Information Officers Council

Version 1.0

Table of Contents

Table of Tables

Table of Figures

1. Data.gov Strategic Intent

In January of 2009 President Obama *instructed* the Director of the Office of Management and Budget (OMB) to issue an Open Government Directive. The resulting *Directive* lays out the specific steps executive departments and agencies must take to implement the principles of transparency, participation, and collaboration outlined in the President's January memorandum. Increased transparency requires a change in the culture of information dissemination, institutionalizing a preference for making Federal data more widely available in more accessible formats. Of particular interest in the context of the Open Government Directive is improving access to data that increases public understanding of Federal agencies and their operations, advances the missions of Federal agencies, creates economic opportunity, and increases transparency, accountability, and responsiveness across the Federal Government. *Data.gov* was launched May 21, 2009 as a flagship Open Government initiative, designed to facilitate access to Federal data. Data.gov was established by the Federal Chief Information Officers' Council and the E-Government and Information Technology Office at the Office of Management and Budget; the General Services Administration operates Data.gov.

The goal of Data.gov is to provide the public with free and easy access to high value, machine readable data sets generated and hosted by the federal government. It will enable the public to easily find, access, understand, and use data that are generated by the Federal government. For data sets that are already available, Data.gov emphasizes making it easier for the public to find and discover data in more usable formats. For data not widely available to the public in the past, the focus is on providing more data more quickly while still protecting and promoting privacy, confidentiality, and security.

This *Concept of Operations* has been influenced by public feedback received through Data.gov, as well as through Federal agency and public reviews of earlier releases of this

Data.gov will be the citizen's front row seat to free and open access to federal data. Data.gov can help realize the vision of open and transparent government by "democratizing" data; making it more available in "Internet time" from anywhere, at any time, readily and reliably.

In his Inaugural address, the President said "Information maintained by the Federal government is a national asset". Citizens from across the nation will be able to download federal data for business, research, analysis and their own pursuits of knowledge. They can combine federal data with their own information and easily share that with others.

The emphasis on access is not new; many agencies already have successful approaches in place. Data.gov continues, reinforces, and focuses efforts at the U.S. government level. This focus is consistent with implementation of OMB's *M-06-02 Memorandum "Improving Public Access to and Dissemination of Government Information and Using the Federal Enterprise Architecture Data Reference Model"*.

On an operational level, Data.gov is continually enhancing service, applications, and policies that ease access to and usability of government data. As an initial matter, the focus has been on creating the public-facing website, as well as less visible elements designed to catalog Federal datasets, improve search capabilities, and publish information designed to allow the end user to determine the fitness for use of published datasets. The goal is to create an environment that delivers improved access to government information, fosters innovation, enhances transparency of government functions, and drives improvements in Agency performance through enhanced information management practices. Realizing these goals requires agencies to:

- Make relevant and informative data and related tools available through Data.gov;

- Help support use and innovation by allowing the public access to data; and
- Design a shared performance management framework centering on high-quality, secure, public information

The purpose of this document is to lay out the overall strategic intent, operational overview ("as is"), future conceptual architecture ("to be"), and next steps. This document is intended to help organize and transform government operations and guide technology development.

1.1. Data.gov Principles

A vibrant democracy depends on straightforward access to high-quality data and tools. Data.gov's vision is to provide improved public access to high-quality government data and tools. Data.gov will provide developers, researchers, businesses, and the general public with authoritative Federal data that are actively managed by data stewards in a framework that allows easy discovery and access.

Key Data.gov principles were established prior to the initial launch and have been refined since the launch. The key principles now include:

The principles behind Data.gov have been refined based on feedback from the public.

1. ***Focus on Access***

   Data.gov is designed to increase access to authoritative sources of Federal data. The goal is to create a transparent, collaborative, and participatory platform that fosters the development of innovative applications (e.g. visualizations and mash-ups) and analyses 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 uses a modular architecture with application programming interfaces (API)[1] *to* facilitate shared services[2] for agencies and enable the development of third party tools. The architecture, APIs, and services will evolve based on public and Federal agency input.

The application development community has provided consistent and strong feedback encouraging an open platform.

[1] An API is a set of instructions and standards that provides a way for software programs to interact with each other. It allows seamless connection between software programs. An API is generally developed by the programmer providing software so that other programmers can make their own software or website more powerful by integrating several programs together. The user of a website or system does not see an API—it works behind the scenes.

[2] A shared service is one that many people can use and is provided from a single source. These are often integrated in other programs or websites in such a way that it seems it is all one service.

3. ***Disaggregation of Data***

   Data.gov promotes and facilitates the disaggregation of data from Federal reports, tools, or visualizations, thereby enabling users to directly analyze the underlying information. Agencies should report data at the lowest analytical unit possible; summaries should be avoided. Data catalogs and tools may additionally combine and display data within a meaningful context.

4. ***Grow and Improve Through User Feedback***

   Data.gov uses feedback from the public to identify and characterize high value data sets, set priorities for integration of new and existing datasets and Agency-provided applications, and drive priorities and plans to improve the usability of disseminated data and applications.

   Citizen engagement is quickly evolving based on the public's uses and experiences with these capabilities.

5. ***Program Responsibility***

   Data.gov is structured to ensure that Federal program executives and data stewards retain responsibility for ensuring information quality, providing context and meaning for data, protecting privacy, and assuring information security. Agencies are 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***

   Data.gov provides the vehicle for agencies to achieve the Open Government Directive mandate to rapidly disseminate new data, as well as immediately improve access to and usability of currently available data. Agencies should ensure that both new and currently available data have sufficient documentation to allow the public to determine fitness for use in the targeted context.

7. ***Embrace, Scale, and Drive Best Practices***

   Data.gov continually implements, enhances, and propagates best practices for data and information management, sharing, and dissemination across agencies, with our international, state, local, and tribal partners.

1.2. Value proposition to the Public

The Administration has emphasized that the three principles of transparency, participation, and collaboration form the cornerstone of an open government. In particular, "Transparency promotes accountability by providing the public with information about what the government is doing. Participation allows members of the public to contribute ideas and expertise so that their government can make policies with the benefit of information that is widely dispersed in society. Collaboration improves the effectiveness of government by encouraging partnerships and cooperation within the Federal government, across levels of government, and between the government and private institutions.[3]" In this section we highlight the role of Data.gov in building these cornerstones.

[3] *http://www.whitehouse.gov/omb/assets/memoranda_2010/m10-06.pdf*

1.2.1. Transparency: Providing Access and Driving Accountability

Providing Access

At the core of Data.gov is making Federal data more accessible and usable to the public. Increasing the ability of the public to discover, understand, and use the stores of government data to increase government accountability and unlock additional economic and social value. Dissemination of public domain data has always been an integral mission activity of the Executive Branch. Specifically, OMB Memorandum M-06-02 requires agencies to "organize and categorize [Federal agency] information intended for public access" and "make [data] searchable across agencies". Data.gov is a major mechanism by which agencies can fulfill requirements in OMB Memorandum M-06-02 in a more consistent and citizen-friendly way. Data.gov takes this traditional activity to the next step by providing coordinated and cohesive cross-agency access to data and tools via a non-Agency-specific delivery channel.

Data.gov is increasingly enhancing the public's ability to find information by offering metadata catalogs integrated across agencies, thus transcending agency stove pipes. A more consolidated source for data and tool discovery allows the public to navigate the Federal sector data holdings without having to know, in advance, how either the Federal government as a whole or an individual Agency is organized.

Based on feedback from the public, Data.gov is working with *USASearch.gov* to develop search capabilities that will further enhance the public's ability to find data, tools, and related Federal web pages regardless of whether they are available on Data.gov[4].

> Easier and more effective search was a commonly requested improvement from the public.

Data.gov also places a strong emphasis on the dissemination of public information generated by the Federal government in platform-independent, standards-based formats that promote creative analysis – via data that are authoritative and granular.

Furthermore, as technology moves from having "data on the web" to a "web of interoperable data," Data.gov will have to evolve to become compatible. Thought is being given to how the development and adoption of semantic web protocols that encode meaning of data in such a way that it is directly interpretable by computers. Through the data web, data aggregation and analyses might be done directly through machine interaction, and new applications and services might be more efficiently created.

Driving Government Accountability

Data.gov has the potential to generate public value by driving better governance through greater accountability, effectiveness, and efficiency in Federal government operations.

> ***Veterans Benefits:*** ***Geographic Distribution of Expenses – Veterans Affairs***
>
> This is a state-by-state and Congressional district-level breakout of Federal expenditures on services to support veterans across America. This dataset provides Federal accountability by making public the details of where Federal expenditures on benefits to veterans go.

Agencies have been asked to post datasets on Data.gov[5] that increase government

[4] Existing policies surrounding the public dissemination of information will be adhered to and extended, if necessary. Inter-Agency working groups populated by data stewards and IT data architects will be significant stakeholders in ensuring consistent implementation of these policies as well as extending guidance, as necessary.

[5] The *Open Government Directive* specifically refers to high-value datasets, which are those that increase public understanding of Federal agencies and their operations, advance the missions of Federal agencies, create

accountability by revealing the results and characteristics of government services to citizens; the public's use of government services; the distribution of funds from the government; and demonstrable results from Federal programs are crucial elements of accountability.

Agencies have been asked to post datasets to Data.gov that can be used to increase government efficiency and effectiveness. Datasets that release government information about how Federal agencies conduct financial management, human resources issues, or other topics in the management of government resources can lead to proposals for improvements to policies and practices.

USAspending.gov

Transparency drives accountability in the example of USAspending.gov, which reveals the recipients and details of over $1 trillion of Federal contracts, grants, and other assistance in a standardized way. Better data quality and more complete details of the procurement process across Federal agencies support efforts to improve the efficiency and effectiveness of government.

Public access to the data underlying *USAspending* has already resulted in several third party (i.e., nongovernment) sites. Figure 1 highlights such a site.

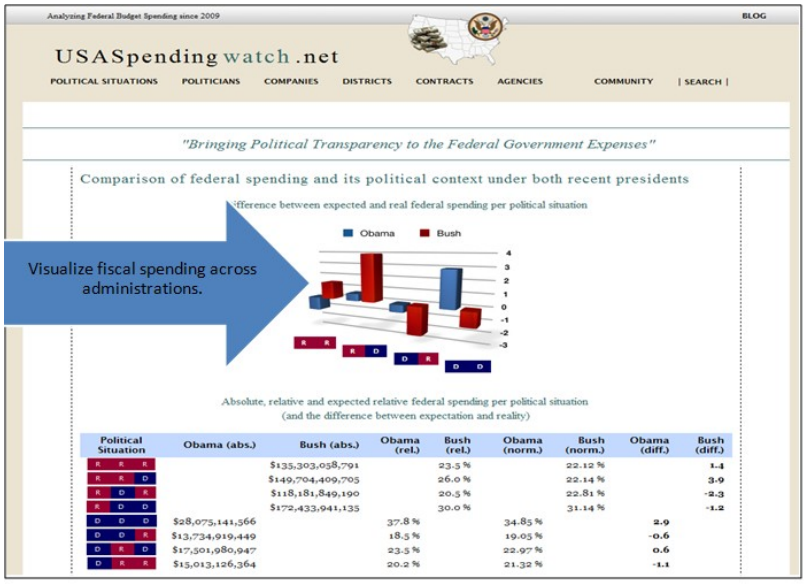

Figure 1: Third Party Participatory Site[6]

1.2.2. Participation: Facilitating Public Education, Engagement, and Innovation

Public participation is a key pillar of open government and is critical to the success of Data.gov. The Administration is particularly interested in developing ways for the public to become more engaged in the governance of our Nation. By increasing opportunities for the public to analyze data, perform research, and build applications, Data.gov provides the resources to spur the creative use of Federal information beyond the walls of government. Education may come in the form of combining Federal and other data to gain new insights into efficiency and effectiveness of government. Innovation might come from the development of web applications, perhaps spurring new economic and socially based ventures.

---

economic opportunity, and increase transparency, accountability, and responsiveness across the Federal Government.

[6] The inclusion of this screenshot, or mention of the site in this document, is not an endorsement of the site.

Specifically, Data.gov supports:

- **Market functioning:** Market participant accountability can be encouraged by providing access to data that describes the behavior and attributes of non-governmental industries and sectors such as capital markets, food and drug industries, public utilities, and transportation systems. Disclosure of government data on private sector activity has the potential to improve corporate behavior because the consequences of their decisions are more publicly available.

  Centers for Medicare and Medicaid Services: Part B National Summary Data File--Health and Human Services

  This dataset provides detailed breakdowns of volume of physician services delivered to Medicare beneficiaries and payments for those services. This helps enable information-centric markets by improving doctors', pharmaceutical developers', care providers', and other healthcare professionals' understanding of the distribution and characteristics of these services across the U.S. These data can be used to look at patterns of Medicare spending and analyze the types of services delivered to address the health needs of the Medicare population.

  The efficiency of information-centric markets benefits directly by ensuring producers and consumers have the maximum amount of information available about their products and purchase decisions. Better information drives innovation, value, and decisions in today's economy. For instance, when the Department of Agriculture makes nutrition information available, families can make smarter eating choices; when the Department of Education makes key information available about colleges and universities, students can make better-informed choices about the quality and cost of education; and when the Department of Labor makes safety information available, employers can better protect workers.

- **Innovators and entrepreneurs**: Success in today's economy is achieved through innovative ideas, sound decision-making, and the ability to adapt to a rapidly changing economic environment. The most important resource for supporting these skills is access and use of valuable data and information. Data.gov provides the opportunity for sophisticated analytical tools to assist with public innovation and decision-making. For example, by providing them rich datasets that will let them grow their businesses, develop new expertise, open new markets, and create jobs. Just as opening GPS data to the world created new markets in mapping, navigation devices, and many other areas, so too other government data is expected to drive innovation in ways we cannot predict. Access to rich data formats, semantic tools and capabilities, and training opportunities has already provided new applications. Examples of some of these can be found at *http://www.data.gov/developers/showcase*. Contests for development of citizen- and business-focused applications will help to focus the efforts of this diverse group from across the world. Specific online discussions forums will help to connect developers to troubleshoot designs, share ideas, and continue to challenge Data.gov to provide more and richer data for this community.

- **Citizen empowerment:** Some datasets and the dialog that emerges around them can lead to citizen empowerment through public participation. Each visitor to Data.gov may find different value in each dataset, discovering answers to their own questions or generating a resource for others.

  Freedom of Information Act Reports

  Multiple agencies, including DOD and the Social Security Administration, have released details from their handling of FOIA requests. These datasets offer details that document the development of a dialogue between the public and the Federal government, thereby supporting the connected citizen and public participation.

An example of the way in which various Executive Branch and external datasets can be combined to empower citizens is application created by Forbes Magazine before the inception of Data.gov to develop their list of *America's safest cities*. This description of the data and how they use it is provided on their website:

> *"To determine our list of America's safest cities, we looked at the country's 40 largest* metropolitan statistical areas *across four categories of danger. We considered 2008 workplace death rates from the* ***Bureau of Labor Statistics****; 2008 traffic death rates from the* ***National Highway Traffic Safety Administration****; and natural disaster risk, using rankings from green living site* SustainLane.com*. It devised its rankings by collecting historical data on hurricanes, major flooding, catastrophic hail, tornado super-outbreaks, and earthquakes from government agencies including the* ***National Oceanic and Atmospheric Administration****, the* ***United States Geological Survey****, the* ***Department of Homeland Security****, the* ***Federal Emergency Management Agency*** *and private outfit Risk Management Solutions. We also looked at violent crime rates from the* ***FBI's*** *2008 uniform crime report. The violent crime category is composed of four offenses: murder and non-negligent manslaughter, forcible rape, robbery and aggravated assault. In cases where the FBI report included incomplete data on a given metro area, we used estimates from Sperling's* Best Places*."*

At some point in the future Data.gov may be able to provide mash-up capabilities that combine data from already existing and well-developed tools and data. Consider, for instance, combining data from the Department of Health and Human Services' Centers for Medicare and Medicaid Services (CMS) Dashboard with data from other sources, such as unemployment statistics from the Department of Labor. Such a data mash-up may be of interest to academic researchers analyzing employment and insurance trends and potentially lead to policies that more effectively address the needs of the American public.

Performance and Accountability Reports

One example which touches on each of these features is the PAR published by each Agency. Historically, this has been a document-centric report of value to a relatively limited audience of students of government performance. The reports are not standardized and for the most part the underlying data is programmatically inaccessible – making it difficult and effort intensive to do additional analysis on the provided information, much less look at cross-Agency trends and performance. In the future, standard reports, such as the PAR, could separate and publish via Data.gov the underlying data. This vision of unbundling the finished report from the underlying data, and potentially augmenting or replacing the traditional document-centric report with data visualizations or web applications, can be extended across many other classes of Government reports.

- **Educators and students:** enabling rich learning opportunities about data, information, applications, and analyses from elementary through graduate school is an important outcome of increasing access to government data. Teachers might use the data in lesson plans or to develop online games, virtual events, and in-class exercises can be created, facilitating the understanding of what data are how they can be used. Older students might use Data.gov to compete in their school science fairs. Undergraduate and graduate students might submit apps and mash-ups, compete in contests, and compete for internships.

1.2.3. Collaboration: Feedback and Outreach

Collaboration, communication, and outreach with both the Agencies who provide our data and the public, who are the ultimate customers, are critical to the success of Data.gov. The Data.gov solicits

feedback and collaboration with citizens and developers through social media like Facebook, Twitter, and the Data.gov weblog. Internal to the Executive Branch, Data.gov engages in outreach and training for agency data stewards and other points of contact, including Open Government representatives. Finally, Data.gov has shared experiences with and provided assistance to state, local, tribal, international and non-governmental partners.

Feedback

Data.gov is not just about disseminating information but is also about interacting with the public. Data.gov provides agencies with a new and important mechanism for understanding the perceived value of their datasets by public users. Thus far, members of the public can provide feedback on published data and tools via the Data.gov site; through Open Government discussions; through the Data.gov forum (IdeaScale); and through social media such as Twitter (@usdatagov) and eventually through Facebook. As Data.gov continues to evolve, the public and government agencies will be provided with new and more robust ways to obtain feedback directly from the end users of their authoritative data

Data.gov has solicited and received feedback from users from four main venues:

1. **Data.gov website** provides members of the public with several opportunities to collaborate in the development and evolution of Data.gov. They can provide specific narrative feedback on published data and tools, rate and comment on the value and quality of current datasets, nominate new datasets to add to Data.gov, and suggest ways to improve the site overall. For datasets already suggested by the public through direct feedback, there is a "user suggested dataset" icon embedded next to each dataset.

2. **IdeaScale** is a technology service that allows registered users to submit ideas around a particular topic (in this case, Data.gov) and organize them under particular topics (e.g., solution architecture, Agency next steps). Other users can review the ideas, provide comments, and "vote" ideas up and down based on interest. Ideas with the most interest filter to the top, while ideas with the least interest fall to the bottom of the site. Users can also choose to track an idea and be notified if there are new comments in the discussion (through email or an RSS feed).

   The IdeaScale forum is accessible directly through the data.gov website's Dialogue tab (also available via *http://datagov.IdeaScale.com*) where users can provide feedback and suggest datasets and new features. As of June 2010, over 500 people have joined the Data.gov IdeaScale site and have posted 152 ideas. The users of the site posted 351 comments and cast 2404 votes. Users are continuing to vote and post comments, but the largest time of activity was in mid-December 2009 and again in mid-January 2010.

   Specific ideas and suggestions that have been submitted to the Data.gov IdeaScale site are highlighted throughout this document. Among the comments received were suggestions related to website look and feel, additional features, improved search and categorization, and requests for more datasets.

3. The **Recovery Dialogue on Information Technology Solutions**, held in May 2009, was a week-long, online Recovery Dialogue on IT Solutions that assembled a broad community of vendors, thinkers, and consumers in the IT arena to answer a central question: What ideas, tools, and approaches can make Recovery.gov a place where the public can monitor the expenditure and use of recovery funds? The Dialogue's sponsor, the Recovery Accountability and Transparency Board, identified five discrete areas in which they expected the Dialogue's participants to potentially provide innovative solutions needed to fulfill their mandate: data collection, data storage and warehousing, data analysis and visualization, website design, and waste/fraud and

abuse detection. While the focus of the Dialogue was Recovery.gov, the Dialogue associated provided Data.gov with access to a wealth of ideas for addressing these challenges. Particularly relevant were discussion regarding the need for data collection solutions to address the large variety of reporting systems, software/hardware platforms, data formats, and issues regarding the use of legacy systems. Furthermore, since the data will be coming from disparate areas in different formats, it will be important to collect and rationalize it in order to provide transparency and make it available for analysis by both government and the public. Quality assurance and the integrity of information are paramount. Such sites as Recovery.gov, USAspending.gov, and Data.gov demand an intuitive public interface, able to present the complex interrelationships among data sources.

4. **Feedback on the draft versions of the CONOPS**, submitted through IdeaScale. The many comments and suggestions provided by the public have been incorporated into this document, reflecting the open and collaborative nature of Data.gov.

The core users of Data.gov currently include the following groups: members of the public with a general interest in certain government programs, technical developers, visualization experts, oversight organizations, teachers, students, researchers, academic researchers, businesses, media, mission advocates, application developers, public sector employees, and individuals interested in knowledge discovery. These groups are identified in Table 1, along with the general use cases they would each have for Data.gov.

Table 1: Core Users of Data.gov [7]

Core Users	Use	Avenues for Interaction
General Public	The general public can use the platform to download datasets. The general public can also discover and access Federal data via third-party visualizations, applications, tools or data infrastructure.	Website, Tools (Agency-Provided and Third Party)
Businesses	Determine novel investment opportunities or underserved markets based off of data available on Data.gov.	APIs, Website, Tools
Application Developers	Application developers can develop and deliver applications by leveraging the raw data, APIs or other methods of data delivery.	APIs, Third-Party Data Infrastructure, online training
Government Mission Owners	Mission owners can expand access to and leverage data from their public sector partners to enhance service delivery, drive performance outcomes and effectively manage government resources.	Website, Tools (Agency-Endorsed)

[7] This is not an exhaustive list and is used for illustrative purpose only. Access modes are both direct and indirect. Direct access is discovering and using the data directly from Data.gov and Agency web sites. Indirect access happens when the public accesses third-party applications, visualizations, or data infrastructure tools that in turn access Federal data via application programming interfaces (APIs) or bulk download of datasets.

Core Users	Use	Avenues for Interaction
Data Infrastructure Developers	Data infrastructure developers can increase the utility of Data.gov by enhancing its search capability, metadata catalog processes, data interoperability and ongoing evolution.	APIs
Research Community	The research community can help unlock the value of multiple datasets by providing insight on a plethora of research topics.	Website, APIs
Data Infrastructure Innovators	Existing entities and new ventures developing innovative data and application offerings that combine public sector data with their own data.	Website, APIs, Bulk Downloads

The user groups are all encouraged to continue providing direct and indirect feedback to Data.gov. Direct feedback takes the form of emails, comments, and ratings posted through the website. Indirect feedback includes blog postings, tweets, social media sites, magazine articles, conference panel discussions, and both traditional and new approaches to engaging in the Data.gov "conversation". A combination of tools exists to collect and utilize indirect feedback, and the Data.gov team welcomes ongoing suggests of other avenues of communication that should be engaged. Feedback has influenced the contents of this concept of operations as well as updates that have been made to Data.gov since its launch. These groups are integrated into the concept of community participation outlined above. They are noted separately here in that some users of the site may or may not participate in a broader community.

Outreach

Data.gov is in the process of building active collaborations with open government organizations; application developers; and state, local, tribal, and international governments. Below we describe each of these.

State and Local Governments

Although Data.gov does not catalog state, local, or tribal datasets, there is a shared benefit of cross-promoting efforts to catalog and make non-Federal data assets more transparent. State, local, and tribal governments are encouraged to leverage the thoughts, ideas, and patterns used by Data.gov to develop their own Data.gov style solutions, and to create views of Data.gov that are appropriate for local needs. State, local, and tribal governments are also encouraged to inform the Data.gov team of their own implementations so that Data.gov can link to those specific sites. The outreach to this community is already underway (see Section 1.2) by Data.gov team member attending meetings and participating in events for these communities.

> Public interest in Data.gov has resulted in many other Data.gov inspired sites around the world, from Canada to the UK to Estonia.

Furthermore, state, local, and tribal governments are encouraged to innovate new and interesting ways of cataloging, presenting, searching, and visualizing their data. As innovations are implemented, non-Federal governments are encouraged to share these breakthroughs with the Data.gov team for potential use on Data.gov.

Data.gov has established an active collaboration with state and local governments. The goal of that collaboration is to provide a platform for sharing their data, linking to their own open data sites, or creating localized views for their neighborhoods. An advisory group representing these groups is being formed to ensure that Data.gov provides access to *all* government data, respecting the differences between levels of government and the needs of the citizens represented by them.

Since the launch of Data.gov in May 2009, many state and local governments have launched their own data catalog sites to better share public sector data. The *map shown* in Figure 2 has the state, local, and tribal areas highlighted where new Data.gov-inspired catalog websites have been launched. Several cities such as *New York* and *San Francisco* have launched Data.gov-inspired websites. Countries such as *New Zealand*, *Australia*, and the United Kingdom have launched sites inspired by or similar to Data.gov.

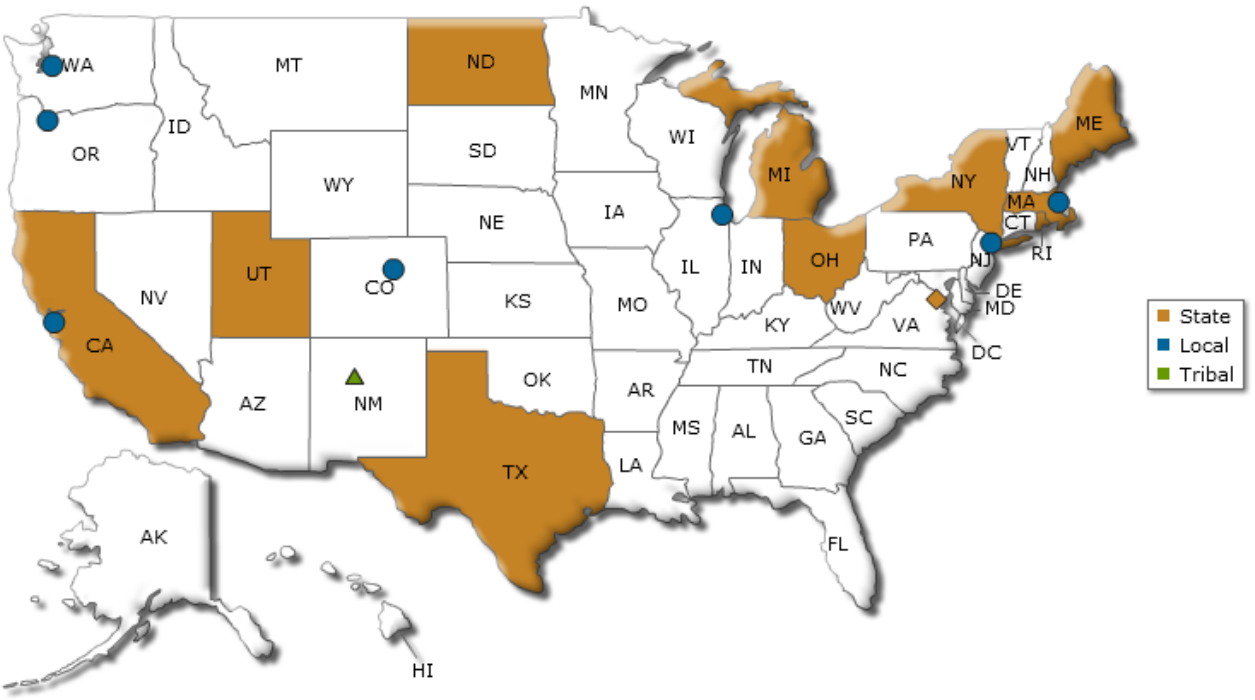

Figure 2: Current (July 2010) State, Local, and Tribal Data Dissemination Sites

International Collaboration

For the international community, Data.gov looking across national boundaries at the data landscape: from data standards and federated ontologies to issues of global importance (such as environmental disasters, climate change, and access to medical care).

Key stakeholders in the development and continual improvement of Data.gov will include the relevant bodies that set international standards related to Data.gov's processes, including the *World Wide Web Consortium*, *International Standardization Organization*, *National Institute of Standards and Technology*, or the actual standards themselves, such as Dublin Core.[8] As Data.gov evolves, it may be necessary to build upon or add to these standards, and even make recommendations to these organizations to update their standards. Data.gov will look to the expert domain communities within the Federal government to work with these organizations and recommend adjustments to Data.gov's processes and metadata requirements.

Expanding Collaboration

As Data.gov looks toward the future, the focus is on the additional communities (e.g., educators and students, developers, international collaborators, and state, local, and tribal governments). Events, forums, contests, virtual science fairs are only some of the ways that Data.gov will seek to engage, energize, and support the efforts of these emergent communities. In all cases, the Data.gov team is committed to not only creating an environment that allows easy discussion with these communities on the Data.gov site, but is also bringing the conversation to the forums, meetings, and settings where these groups already congregate.

[8] This is not an exhaustive list and does not account for all organizations that will be relevant to this issue.

In addition Data.gov may in the future explore extending its functional capabilities to include ranking and rating of the public suggestions for priority datasets and tools. Another capability could include the 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. Beyond these Data.gov-centric collaboration concepts, Data.gov may be a source of data for non-government sites that develop their own forums for collaboration. One such site was developed as part of the Apps for America contest sponsored by The Sunlight Foundation. Figure 3 highlights such a site that not only provides a map of information about earthquakes and storms sourced from Data.gov, but also displays in real time the tweets on natural disasters via the Twitter API and shows the location(s) of the members of the U.S. Congress interested in natural disaster-related problems (using the Capitol Words API).

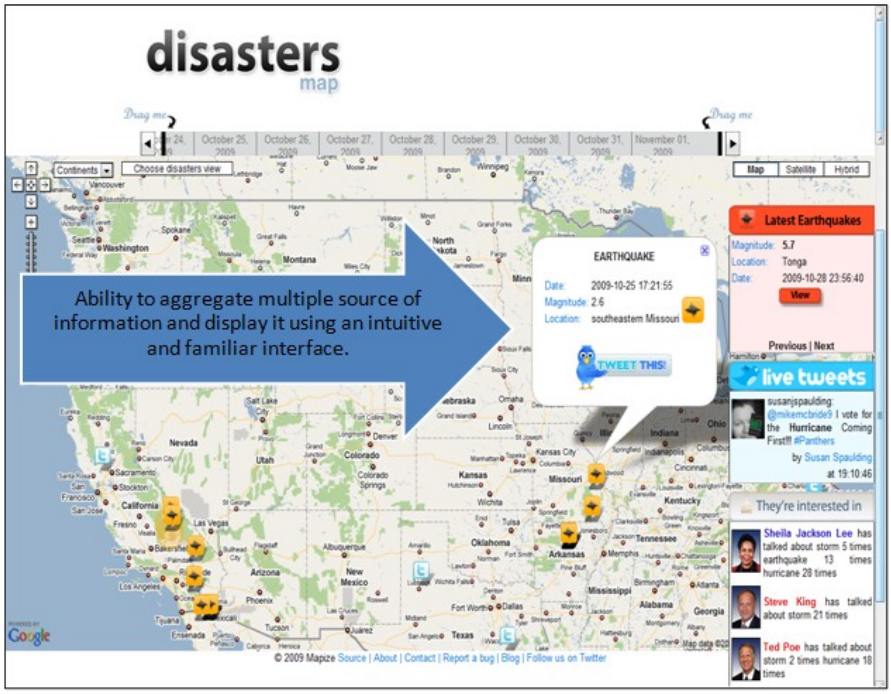

Figure 3: Third Party Disasters Map Mash Up and Collaboration Example[9]

Several comments submitted on the IdeaScale site centered around enhanced feedback and collaboration to improve Data.gov. Generally, participants were interested in more interactivity around specific datasets. For example:

- One user was interested in error correction on datasets, saying, "There should be ways for the public to add notes about possible problems they have found with the data and a feedback mechanism to ensure that others know about these possible problems." Another participant said that they were interested in "tools for crowd-sourced analysis, such as labeling points of interest on a map, interesting pages in a large document, or parts of a dataset that need more attention." The first point has been incorporated into Data.gov's functionality, and the second is under review.

[9] The inclusion of this screenshot, or mention of the site in this document, is not an endorsement of the site.

- Other participants were interested in discussion forums either around specific datasets or for general topics, “Why not create a forum where the citizens have a place to discuss topics that impact us all and maybe the government will monitor to help try and follow the sway of the people and not just the sway of those with the most money?” Such a forum would enable the public to talk what they have done with particular datasets or questions they might have.
- Similar to the above, there was interested from a user for “wish lists” both for analyses that the public might want for a particular dataset and for datasets that are not yet available.
- One user noted that, instead of general dialogues with the public about datasets, it would be useful to have “a means whereby direct dialogue could be established between the public and actual SMEs [subject-matter experts] or communities of SMEs, associated with particular datasets.” The participant acknowledged that considerations of privacy, security, and time were important to consider. This idea may be addressed by the Data.gov/Communities (e.g. Data.gov/Education and Data.gov/Health) currently in development.
- On the developer end, there was interest in a developer wiki to encourage collaboration and sharing of tools, sample code, and sandboxes amongst Data.gov developers, developers at government agencies, and the general community of interest. Participants indicated that it would be useful for adoption of particular APIs and using the data. Another user also asked that, for each dataset, there be “a set of developer-general recipes for data manipulation so that future developers can stand on the shoulders of previous users of this dataset by sharing tools and techniques.” Data.gov plans to incorporate these suggestions as the Developers’ Corner page matures.
- Some participants indicated that additional dialogue around the intent and use of Data.gov is needed to the public in general and to specific groups. One idea was for Data.gov to open a dialogue with the U.S. Congress to “ensure that Data.gov addresses the data needs of these oversight committees so that Senators and Congressmen alike can make better informed decisions that ultimately affect Agency responsibilities, staffing, performance expectations, and funding.” A commenter on this idea said that the idea should go further and that Data.gov be the only means of providing information to Congress and other government agencies.
- One participant said that Data.gov should be talking with the public more to find out what data the public wants or needs to know about. The idea this participant submitted was that, “the Data.gov effort begin with a dialogue of the 'public' they envision using the data feeds on Data.gov.” The user provided a few topics that this dialogue should focus on, including knowing what issues about federal Agency performance are important and what formats the public would like data. Data.gov’s social media campaign, currently active on Twitter and soon to be on Facebook, will provide the public another avenue to hear from and speak to Data.gov.
- Some users acknowledge that, in order to get many of the collaboration additions that were submitted on the IdeaScale site, particularly around feedback about datasets, changes to Paperwork Reduction Act are needed.
- Outreach focuses on promoting Data.gov, both to encourage wider use amongst the public, as well as to persuade greater participation and involvement from government agencies. A number of the participants discussed this topic, specifically as it related to improving the available data and bringing more information to the site. One user suggested, “Data.gov needs a Data.gov evangelist who can be the community manager out on the road listening and talking and generally spreading the word on Data.gov and its value.” According to this idea, this position

could be focused on working with developers on applications. Data.gov has implemented these ideas and appointed a "Data Evangelist" and a team of communications specialists to lead the outreach effort to both Agencies and the public.

- A simple way to inform those interested in Data.gov would be to publish a roadmap with milestones for Data.gov on the website. The participant who suggested this idea would help "public can see when planned innovations will occur and that they are actually influencing the evolution."
- A common desire of the public is to see the real-world applications of whatever service or data is available. A number of ideas suggested that there be more communication about how Data.gov can help address the issues that the United States will face in the coming decade. There was also a call for highlights of real-world applications making use of the information from Data.gov and for government sponsored contests to create more applications. This suggestion led to the new Apps Showcase, and is influencing the development of the Communities (e.g. Data.gov/RestoretheGulf).

1.3. Value Proposition to Executive Branch

Improving the Federal Data Management Process

Data.gov has the potential to improve the Federal government's return on investments in collecting and managing the data themselves by providing an environment that:

- transcends Agency stovepipes
- encourages data to be disseminated in reusable and interoperable formats
- enhances the ability of government agencies to find data and data sharing opportunities across the Federal space
- facilitates enhanced search abilities across the entire Federal landscape of data

Further, many opportunities exist for adding value such as exploring more timely release of in-process data assets rather than accumulating, processing, and disseminating data on longer, agency-centric timelines. In particular, faster release of data would support more timely, third-party analysis and have the potential to empower more proactive public-initiated dialog.

In the initial stages of opening government data, the Data.gov team focused, necessarily, on the federal government community to be able to populate datasets in the Data.gov system. This community has grown to encompass more than 250 individuals representing agencies across the government. This community has shepherded a transformational change within each Agency—making data visible and changing the way in which data is gathered, formatted, and published. This change is far-reaching and touches upon processes, services, and systems throughout the government and truly has initiated a new way of making information transparent and accessible.

Data.gov Helps EPA

The Environmental Protection Agency (EPA) has a significant data-oriented mission. EPA collects data and then makes them available to users via the public domain. One such category of EPA data is the toxic release inventory (TRI). TRI data are made available through the EPA website and, in July 2009, were integrated into Data.gov. After the TRI data were featured on Data.gov, the data were downloaded more than 1000% more frequently than during July 2008. While still early, the frequency of download continues in excess of 500% greater year-over-year.

Such accessibility also addresses one additional goal of Data.gov: enhancing cross-government data sharing by highlighting access to information held by one Agency that might be of value for achieving

another Agency's mission. Data.gov will eventually provide the opportunity for agencies to leverage Data.gov shared data storage services if they so desire.

Since most agencies have information dissemination as part of their mission, Data.gov is a key component for improved mission delivery. It is a delivery channel to enable agencies to make their data more accessible, discoverable, understandable, and usable.

Delivering value through making data transparent is an important aspect of achieving core agency missions.

Agencies that target their contributions to Data.gov in order to expand their offerings in each of the public categories will make better use of Data.gov than agencies that simply release a large number of datasets that do not more clearly support public value. As agencies focus on datasets that address these categories, the public will benefit from higher value datasets and more valuable resources.

Data.gov promotes government-wide data-management practice improvements. Agencies may choose to use Data.gov as their primary means of dataset information dissemination to the public and forgo the need to maintain potentially redundant processes and infrastructure for publishing their data. Specifically, agencies can use Data.gov not only to store their metadata via the Data.gov metadata storage shared service, but agencies can also forego management of their own data storage infrastructure by leveraging a data storage shared service described in Chapter 3.

As additional data and tools are made available through Data.gov, the value of Data.gov will compound its value to both agencies and the public. As improvements are made to metadata, data quality, search, discovery, and access tools, Data.gov will become an important resource to user groups, leading in turn to greater visibility and use of data. As the value to the public increases, so too will the value to the agencies supporting the public, which in turn will result in increased attention on providing datasets that provide additional value. In this manner, agencies have a vested interested in not only their own active participation, but in the active participation of their peer agencies.

From an individual Agency perspective, Data.gov is another vehicle for providing information and outreach about the value of the missions they serve. As illustrated by an example in Figure 4, Data.gov includes revolving panels of "featured tools and datasets" that provide high profile visibility showcasing high-quality data and tools provided by the Federal government. These featured tools and datasets rotate to provide the regular visitor to Data.gov a new perspective on each visit. In the future, datasets that relate to current events, disaster or emergency management, or anticipated interest areas will be featured in this area. As an example, datasets providing earthquake data can be moved to this more prominent location during an earthquake event to facilitate access to a dataset that is anticipated to be in high demand. This aspect further enhances the service value of Data.gov and the datasets it provides access to.

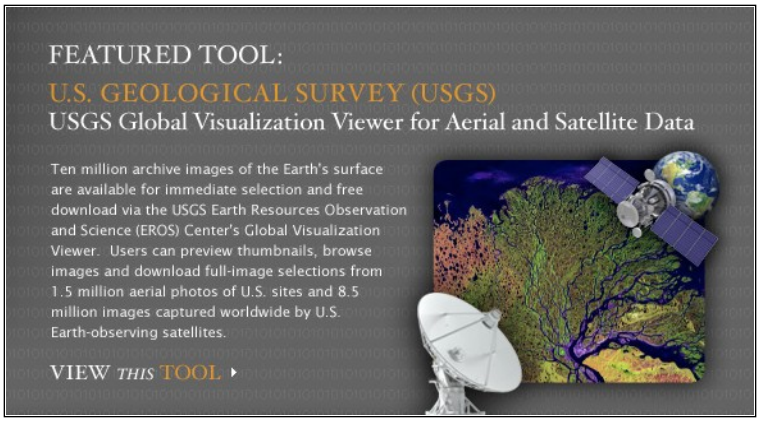

Figure 4: DOI (USGS) Featured Tool Panel Example

Agencies that actively participate in Data.gov not only share their data more widely, but also increase the public's awareness of their works in key mission areas. Active participation in Data.gov increases overall visibility and can engender a greater trust and appreciation for Agency missions, their roles, and their overall performance in the service of the country. Transparency of Agency data provides the public with the ability—either through government tools, third-party web applications, or other means—to understand their government, its impact on their lives, and hold it accountable. This transparency can also translate into the discovery and implementation of collaborative initiatives with other Federal organizations.

1.4. Measuring Success

Data availability, data usage, and data usability are key to successful implementation of Data.gov. Primary and secondary metrics for measuring availability and for measuring success are described below.

Through direct feedback, blog postings, news articles, and conference speakers, the public has provided much feedback on how to measure the success of data transparency.

1. Data Availability

Agency participation will be evaluated based upon the quantity of data that they make available through Data.gov, relative to the total amount of data that the Agency has eligible for such access (i.e., the release of which does not compromise privacy, confidentiality, security, or other policy concerns).

Agencies should prioritize information dissemination efforts to accelerate dissemination of high-value datasets in accord with mission imperatives, agency strategy, and public demand. Data.gov is designed to facilitate access to Federal datasets that increase public understanding of Federal agencies and their operations, advance the missions of Federal agencies, create economic opportunity, and increase transparency, accountability, and responsiveness across the Federal Government – i.e., "high value" datasets.

In regards to the Government Performance Results Act (GPRA), agencies may use the dissemination-related metrics present in this *Concept of Operations*.

Table 2: Availability Metrics

Metric Title	Definition	Where Posted
Website Availability	The percentage of time the website is functioning without errors	DMS
Number of Tools	The number of technical tools on Data.gov	DMS
Response Time Objective (RTO) and Recovery Point Objective (RPO)	Amount of response time from the website failing to the issues being resolved	DMS
Federal Agency Participation at Data.gov	What the agencies contribute	DMS and Data.gov
Number of POCs	Number of Agency POCs that have completed the Data.gov Orientation and the Data Management System(DMS) trainings at least once	DMS
Number of Trained POCs and Data Stewards	Number of Data Stewards that have completed the Data.gov Orientation and the Data Management System(DMS) trainings at least once	DMS
Number of Published Datasets	Number of datasets that have been published on the Data.gov website after completing all steps in the approval process.	DMS
Application Programming Interfaces (APIs) uptime	The percentage of time the APIs supplies by Data.gov are functioning without errors	DMS
Public Web Accessibility Metric	This metric measures the overall (aggregate) percentage of crawled public web pages that are inaccessible due to technical search related errors. Technical search related errors (HTTP Status Codes) are categorized by the most common type of error (see below). Uncommon technical errors will be categorized as other. A sub-measure is defined for each of the six technical error categories to aid in the identification of recurring technical errors as well as help IT staff to resolve the error.	DMS
Response Time Efficiency Percentage	The average response time to search individual geo-spatial collection records, calculated by measuring the average time taken to search a web index.	DMS
Number of Geospatial Data	GeoData is a catalog of geospatial information containing thousands of metadata records (information about the data) and links to live maps, features, and catalog services, downloadable datasets, images, and map files. The metadata records were submitted to the portal by government agencies.	GeoData.gov

2. **Data Usage**

Usage metrics include the number of Data.gov page views, downloads of data, API calls, success in facilitating innovation as demonstrated through the number and scope of third party applications, feedback from users and intra-governmental collaboration on Data.gov-related content.

Success in facilitating innovation could be measured through proxies including the diversity of use of Data.gov's content and Data.gov's propagation. Diversity of use will be ranked by the number of third-party applications, references to Data.gov and its datasets in publications and analyses, and use of the third-party applications. User feedback metrics could measure the volume and sentiment of feedback, both overall and on a dataset basis.

Table 3: Usage Metrics

Metric Title	Definition	Where Posted
Application Programming Interfaces (APIs) calls per dataset	The number of times any external applications searches against standard or optional parameters of the provided Data.gov API and returns that dataset information	DMS
Daily Visitor Statistics by Month	Illustrates maximum and minimum daily visitors to Data.gov. Also includes daily average and monthly total visitors.	Data.gov
Monthly Visitor Statistics	This chart includes the total number of visitors to Data.gov and the year-to-date total number of visitors.	Data.gov
Monthly Downloads by dataset category	These numbers represent the number of times a user has clicked on the "XML" or "CSV" (for example) links in the Raw Data Catalogs to download datasets and user downloads of tools in the Tool Catalog available in these categories.	Data.gov
Monthly Download Trends	These numbers represent the number of times a user has clicked on the "XML" or "CSV" (for example) links in the Raw Data Catalogs to download datasets and user downloads of tools in the Tool Catalog available in these categories. Also includes year to date number of downloads	Data.gov
Monthly number of website hits	This chart represents year to date number of hits on Data.gov.	Data.gov
Top 10 Visitors by Country (by Month)	Top 10 visitors identified by country monthly basis	Data.gov
Top 10 Visitors by State (by Month)	Top 10 visitors identified by state on a monthly basis	Data.gov
Top 10 most Downloaded Datasets (All Time)	Top 10 datasets since the inception of Data.gov that have been identified as downloaded	Data.gov
Top 10 most Downloaded Datasets in the Last 30 Days	Top 10 datasets in the last 30 days that have been identified as downloaded	Data.gov
Top 10 highest ranked Datasets	Data.gov utilizes the Bayesian Rating (BR) to determine which datasets are the highest ranked. The Bayesian Rating uses the Bayesian Average. This is a mathematical term that calculates a rating of an item based on the "believability" of the votes. The greater the certainty based on the number of votes, the more the Bayesian rating approximates the plain, unweighted rating. When there are very few votes, the Bayesian rating of a dataset will be closer to the average ratings of all datasets that were voted on.	Data.gov

3. Data Usability

Usability will be measured by how clearly and completely the strengths and weaknesses of Agency data are conveyed through technical documentation. This will be measured via completeness of the structured data provided and maintained by agencies that represent integrated datasets in the Data.gov platform. Additional metrics include: how detailed are the key words that feed the search; the degree to which semantic web approaches as described in Chapter 3 are used; proper registration of APIs; proper descriptions for all relevant columns in a dataset; and user dataset scoring.

Table 4: Usability Metrics

Metric Title	Definition	Where Posted
Errors per dataset	Number of errors reported for each dataset	DMS
High-value Dataset Discovery Results Efficiency Percentage	A measure of the ability to identify high-value datasets through web crawling. For the purposes of the pilot, a dataset is defined as any geodata or raw data with file format extension types. High-value datasets are those datasets that have specific file format extensions for geodata and/or raw data (e.g., xml, csv, or xls for raw data).	DMS
Meta-data Completeness Percentage	A measure of the completeness of the mandatory geospatial meta-data values contained within the high-value datasets identified and collected through web crawling. Mandatory geospatial meta-data elements are defined by the Federal Geographic Data Committee (FGDC) endorsed standards (FGDC-STD-001-1998). It is calculated by measuring the number of mandatory geospatial meta-data elements within discovered high-value datasets with a corresponding value (i.e., non-blank).	DMS
Match Results Percentage	A measure indicating the percentage of crawled geo-spatial mandatory meta-data elements that match the mandatory meta-data elements in the data.gov metadata template. Required data.gov metadata elements are defined as those that are common for all domains and those that are specific to a domain (e.g., geospatial, statistics).	DMS

2. Data.gov Operational Overview

2.1. Policy governing datasets accessible through Data.gov

Data.gov is designed to provide direct access to Executive Branch datasets and tools for analyzing and visualizing those data. All datasets accessed directly from the Data.gov catalog should be authoritative sources that meet pre-existing statutory mandates and Executive Branch policy for information dissemination. Below we describe policies of particular interest in the context of Data.gov, including those associated with information quality, privacy and confidentiality, computer security, and accessibility. Specific Agency responsibilities for assuring compliance are described in Section 2.5.

2.2. Information Quality

Many Executive branch data collections are subject to the ***Paperwork Reduction Act (PRA[10])*** The PRA was designed to, among other things, "ensure the greatest possible public benefit from and maximize the utility of information created, collected, maintained, used, shared and disseminated by or for the Federal Government" and "improve the quality and use of Federal information to strengthen decision-making, accountability, and openness in Government and society."[11] Federal agencies play a critical role in collecting and managing information in order to promote openness, reduce burdens on the public, increase program efficiency. OMB is required to review information collections subject to the PRA. A central goal of OMB review is to help agencies strike a balance between collecting information necessary to fulfill their statutory missions and guarding against unnecessary or duplicative information that imposes unjustified costs on the American public. In this regard, OIRA evaluates whether the collection of information by the Agency:

- is necessary for the proper performance of the functions of the Agency, including whether the information has practical utility;
- minimizes the Federal information collection burden, with particular emphasis on those individuals and entities most adversely affected; and
- maximizes the practical utility of and public benefit from information collected by or for the Federal Government.

OMB also reviews the extent to which the information collection is consistent with applicable laws, regulations, and policies related to privacy, confidentiality, security, information quality, and statistical standards. In addition, OMB coordinates efforts across Federal agencies in shared areas of interest and expertise. A public inventory of currently approved information collections, including justifications and protocols, is available at *http://www.reginfo.gov/public/do/PRAMain*

Furthermore, the **Information Quality Act** applies to all data directly disseminated from Data.gov. Specifically, in accordance with Section 515 of the *Treasury and General Government Appropriations Act for Fiscal Year 2001 (Public Law 106-554)*, OMB has published guidelines (available at *http://www.whitehouse.gov/omb/assets/omb/fedreg/reproducible2.pdf*) to help agencies ensure and maximize the quality, utility, objectivity, and integrity of the information that they disseminate. In addition, all Federal agencies are required to issue their own implementing guidelines that include administrative mechanisms allowing affected persons to seek and obtain correction of information maintained and disseminated by the Agency that does not comply with the OMB guidelines.

[10] 44 U.S.C. chapter 35; *see* 5 CFR Part 1320

[11] 44 U.S.C. § 3501

The OMB government-wide guidelines impose three core responsibilities on the agencies:

- First, the agencies must embrace a basic standard of "quality" as a performance goal, and agencies must incorporate quality into their information dissemination practices. OMB's guidelines explain that "quality" is an encompassing term comprising utility, integrity, and objectivity. The concept of quality includes consideration of:
    - the usefulness of the information for the intended users, taking into consideration both public and private decision making;
    - the transparency of the method used to generate the information, ensuring that independent analysis of the original or supporting data using identical methods would generate similar analytical results, subject to an acceptable degree of impression or error;
    - the security of information, ensuring that information is protected from unauthorized access or revision, to ensure that the information is not compromised;
    - the context in which the information being disseminated, ensuring that the information is presented in a manner that is accurate, clear, complete, and unbiased; and
    - the accuracy, reliability, and potential for bias in the underlying information, ensuring that the original data and subsequent analysis were generated using sound research and/or statistical methods.
- Second, the agencies must develop information quality assurance procedures that are applied ***before*** information is disseminated. For scientific information, the practice of independent peer review plays an important role in establishing a presumption that is "objective."
- Third, the OMB government-wide guidelines require that each Agency develop an administrative mechanism whereby affected parties can request that agencies correct poor quality information that has been or is being disseminated. Furthermore, if the public is dissatisfied with the initial Agency response to a correction request, an administrative appeal opportunity is provided.

The scope of the OMB's government-wide guidelines is broad. It spans information related to regulatory, statistical, research, and benefits programs. It covers all Federal agencies subject to the Paperwork Reduction Act, including the independent regulatory commissions. OMB's guidelines define "information" as "any communication or representation of knowledge such as facts or data" in any medium. OMB provided a variety of exemptions from the guidelines to protect individuals' privacy and commercial secrets, and to facilitate press releases, third party submissions in public filings, archival records, personal articles by Agency employees, testimony, subpoenas and adjudicative determinations.

OMB recognized that information quality can be costly and encourages agencies to consider the social value of better information in different contexts. OMB's guidelines recognize that some government information may need to meet higher or more specific standards than would apply to other types of government information. OMB's guidelines encourage agencies to weigh the costs and benefits of higher quality information; the more important the information, the higher the quality standards to which it should be held. Information that is most likely to have influence on important public and private sector decisions requires a higher level of quality[12].

[12] Per OMB's government-wide guidelines, "Influential" information is subject to higher standards of quality. "Influential" means that information that the Agency can reasonably determine that dissemination of the information will have ore does have a clear and substantial impact on important public policies or important

Because information disseminated by the Executive Branch is subject to specific statutory mandates, Data.gov cannot provide direct access to datasets disseminated by other branches of the Federal government, state, local, and tribal governments, nongovernmental organizations, the private sector, or even those Executive Branch data that have been reformatted by members of the public. However, Data.gov does provide indirect access to such external datasets and resources through its state, local, and tribal portal[13] as well as through some of its federated search options.

Within the context of Data.gov, this means that each dataset continues to reside on disseminating agency's own web site; the agency retains responsibility for authoritative source of data, including corrections and updates; the agency retains responsibility for protecting personally identifiable information, information quality, and records retention; and each Agency retains responsibility for impacts of Data.gov access on their site.

Ensuring quality of the datasets on Data.gov is very important so that the public will trust and find value in the information. Below are some of the major points IdeaScale users have identified around information quality of Data.gov datasets. Many of these comments have been addressed already, while others remain outstanding. Data.gov continues to work with Federal agencies to improve the quality of the datasets available to the public.

- **Standard Taxonomies**: Many participants noted that there was not enough information associated with datasets to be useful or that the metadata was not clear enough to be understood. The addition of standard taxonomies and ontologies would expand the metadata template to make it clearer to users viewing a record and easier to find for those searching or browsing for data.
- **Data Definitions**: Some participants noted that some of the associated data dictionaries were incomplete or incorrect; while this issue is not specific to Data.gov, necessarily, it is associated with the idea of complete training for improved data quality for government agencies. One user noted that "every column or field of data should have a definition and that should be available on Data.gov or in a standard format with the dataset."
- **Agency Quality Checks**: Participants noted problems with the metadata entered into data records. One user suggested that there should be "data quality controls checklist that is completed by the submitter before publishing the data," and another suggested embedding a workflow function into the data management system to track releasability of datasets.
- **Standard File Formats**: Participants expressed interest in requiring common file formats for the data so that the public can download and open files easily and so that it will be easier to create mash-ups and integrate data across datasets. Currently, users can search on a set of common file formats, but data formats other than these have made their way onto Data.gov.
- **Error Correction**: Users expressed interest in error correction and crowd-source analysis and information on datasets. One user said, "There should be ways for the public to add notes about possible problems they have found with the data and a feedback mechanism to ensure that others know about these possible problems."

---

private sector decisions. Each Agency is authorized to define "influential" in ways appropriate for it given the nature and multiplicity of issues for which the Agency is responsible. With several important exceptions and qualifications (e.g., privacy, intellectual property rights, and other confidentiality protections) influential information should be reproducible by qualified third parties.

[13] *http://www.data.gov/statedatasites*

- **Transparency**: Some users indicated that more transparency about the data on Data.gov would decrease questions and concerns (e.g., when datasets are updated and why, why particular datasets are not available). Making this kind of information transparent could help user feel better about the quality of the information that is available.
- **Data Quality Confidence:** A participant discussed the usage of information from Data.gov in ways that could potentially be misinterpreted. This participant wondered, "Because Data.gov seeks to make raw data available to a broad set of potential users, how will Data.gov address the issue of data quality within the feeds provided through Data.gov? Currently, federal Agency Annual Performance Reports required under the Government Performance and Results Act (GPRA) of 1993 require some assurance of data accuracy of the data reported; will there be a similar process for federal Agency data made accessible through Data.gov?" One first-step suggestion made was for agencies to include a "confidence" indicator in the metadata template about how "good" a particular dataset is.

Data Security and Privacy

Data.gov is committed to upholding Federal privacy and confidentiality protections. The privacy objective of the E-Government Act complements the National Strategy to Secure Cyberspace. As the National Strategy indicates, cyberspace security programs that strengthen protections for privacy and other civil liberties, together with strong privacy policies and practices in the Federal agencies, will ensure that information is handled in a manner that maximizes both privacy and security.

Federal agencies, including Data.gov, must protect an individual's right to privacy when they collect personal information. This is required by the Privacy Act, 5 U.S.C. 552a, and OMB Circular No. A-130, "Management of Federal Information Resources," 61 Fed. Reg. 6428 (Feb. 20, 1996), and supported by the *Principles for Providing and Using Personal Information* published by the Information Infrastructure Task Force on June 6, 1995. Posting a privacy policy helps ensure that individuals have notice and choice about, and thus confidence in, how their personal information is handled when they use the Internet.

Agencies must not post personally identifiable information on Data.gov or in any other way compromise Federal law and policy. All data on Data.gov must conform to all applicable security and privacy requirements including the *Privacy Act of 1974*, the *E-Government Act of 2002*, applicable Federal security standards[14] including NIST 800-39, and other privacy[15] and confidentiality[16] guidance as issued by OMB (See also: *Office of Information and Regulatory Affairs (OIRA) Information Policy*).

Privacy considerations extend beyond the data itself to include the way that Data.gov measures performance and gathers feedback from the site. All feedback provided through Data.gov is anonymous with no tracking or identifier information captured. Furthermore, performance statistics will be gathered as specified within this document. Performance measures will be specifically targeted at macro use statistics without any identification of specific uses of specific data by individuals or groups.

Data.gov's privacy policy is available on the Data.gov website, or *http://www.data.gov/privacypolicy*.

[14] *http://www.whitehouse.gov/omb/inforeg_infopoltech/#cs*

[15] *http://www.whitehouse.gov/omb/inforeg_infopoltech/#pg*

[16] *http://www.whitehouse.gov/omb/inforeg_statpolicy/#stat_conf*

E-Government Act

Section 207(d) of the *E-Government Act of 2002*, requires OMB to issue policies – "(A) requiring that agencies use standards, which are open to the maximum extent feasible, to enable the organization and categorization of Government information: (i) in a way that is searchable electronically, including by searchable identifiers; (ii) in ways that are interoperable across agencies; and (iii) that are, as appropriate, consistent with the provisions under 3602(f)(8) of title 44, United States Code; (B) defining categories of Government information which shall be required to be classified under the standards; and (C) determining priorities and schedules for the initial implementation of the standards by agencies." In particular, OMB's *M-06-02 Memorandum "Improving Public Access to and Dissemination of Government Information and Using the Federal Enterprise Architecture Data Reference Model"* states that agencies "have three new requirements in this area although many of you are already meeting them in part. As outlined below, you must now: A) organize and categorize your information intended for public access, make it searchable across agencies, and describe how you use formal information models to assist your dissemination activities; B) review the performance and results of your information dissemination program and describe the review in your Information Resources Management (IRM) Strategic Plan; and C) publish your IRM Strategic Plan on your public website."

Accessibility

Section 508 of the Rehabilitation Act requires that Federal agencies provide individuals with disabilities who are either Federal employees or members of the public seeking information or services with access to and use of information and data that are comparable to the access to and use of the information and data by such Federal employees or members of the public who are not individuals with disabilities. The Data.gov website is designed and tested accordingly to ensure conformance to the requirements for Section 508.

The commitment to accessibility for all is reflected on this site in our efforts to ensure all functionality and all content are accessible to all Data.gov users. The Data.gov site is routinely tested for compliance with Section 508 of the Rehabilitation Act using a technical standards check-list, in-depth testing with screen readers, policy experts, and persons with disabilities. For more information on Section 508 technical standards please visit ***www.Section508.gov***.

> *The power of the Web is in its universality. Access by everyone regardless of disability is an essential aspect.*
>
> **Tim Berners-Lee, W3C Director and inventor of the World Wide Web**

In addition to maintaining Section 508 compliance, Data.gov is also routinely reviewed for alignment with the latest Web Accessibility Initiative Guidelines for W3C. The Web Accessibility Initiative Guidelines at ***www.W3.org/WAI/*** define how browsers, media players, and other "user agents" support people with disabilities and work assistive technologies.

> "Web accessibility means that people with disabilities can use the Web. More specifically, Web accessibility means that people with disabilities can perceive, understand, navigate, and interact

with the Web, and that they can contribute to the Web. Web accessibility also benefits others, including older people with changing abilities due to aging."[17]

Images on the site contain 'alt tags,' which aid users who listen to the content of the site by using a screen reader, rather than reading the site. Likewise, a 'skip to' link provides these users with a method for bypassing the header and going directly to the main content each time a page is accessed. Text transcripts accompany audio clips, and closed captioning is available on videos.

The Data.gov website is being updated frequently to make it as accessible as possible. Users of assistive technology (such as a screen reader, eye tracking device, voice recognition software, etc.) who have difficulty accessing information on Data.gov, can contact the site and provide the URL (web address) of the material they tried to access, the problem experienced, and their contact information. A Data.gov team member will contact the user and attempt to provide the information directly.

Mosaic Concerns

As we increase the availability of government data to the public in more easily accessible and interoperable formats, we need to be careful to protect national/homeland security and safeguard privacy and confidentiality. As required by law, agencies have in place processes for screening datasets for national/homeland security and privacy/confidentiality, per existing statutes.

- The Open Government Directive requires agencies to re-visit their past decisions not to release datasets to the public. We are working with the agencies to ensure the type of disclosure review that is necessary to reduce the risk of breaches of privacy/confidentiality and/or national/homeland security.
- The content of any individual dataset may not pose a threat to national/homeland security or a risk of a breach in promises of confidentiality or privacy. However, the collection of many datasets disseminated jointly, especially within the context of an environment that encourages 'mash ups,' may increase the risk. Specifically, Data.gov is cognizant of concerns that the cross-government cataloging and searching associated with Data.gov may facilitate identification of relationships that may not otherwise be obvious.

Screening procedures to help reduce the risk associated with mosaic concerns - both from the national/homeland security perspective and from the privacy/confidentiality perspective - have been developed. OMB, NSC, and GSA, with input from other Executive Branch agencies with specialization in these areas, continue working together to further evaluate and enhance Federal data dissemination guidelines to guard against intentional and unintentional unmasking of sensitive or personally identifiable information and/or national/homeland security-sensitive information. As additional opportunities to enhance policies and procedures for evaluating datasets for mosaic concerns are developed, agencies agency training will be provided.

2.3. Determining Fitness for Use and Facilitating Discovery

The term "metadata" means an external description of a data resource. The term is often used to describe information that enables: (1) discovery of data, (2) understanding the provenance and quality of the data, or/and (3) analysis of the data via a set of machine readable instructions that describe the data and its relationships, as might be required by a data web.

[17] http://www.w3.org/WAI/intro/accessibility.php

The Data.gov website uses a metadata template as a means of describing the core attributes of each dataset and data extraction/mining tools cataloged. The information in the metadata template both powers the search engine and provides access to information about the potential utility of the dataset for a given use.

The Data.gov metadata template currently includes descriptive information about the context of the data collection, the study design, dataset completeness, and other factors that might influence a data analyst's determination regarding the utility of the data for a specific purpose. The core elements in the Data.gov metadata template were based on the *Dublin Core Metadata Element Set, Version 1.1 standard*.

Given the use of the metadata template for populating the search database and providing context information about the available datasets, the most critical elements of the metadata template include the data descriptions, keywords, data sources, and URLs for technical documentation. Agencies should think both broadly and specifically when selecting key words – the robustness of the text-based search capability will drive the extent to which users can find the data in which they are interested. An inter-Agency committee of metadata and search experts will continue to refine the *Data.gov* metadata template as both the vision and architecture of the site evolve.

Domain specific communities within the Federal government are encouraged to develop their own supplemental metadata standards that would be harmonized with the core metadata specification and co-exist in a federated context. For instance, in addition to the core elements defined for datasets and tools, the Data.gov metadata template currently accommodates additional metadata elements for datasets that are classified as "statistical". On the other hand, for communities that already have their data assets cataloged using a standard metadata format, APIs can be built to "mine" that metadata for the elements needed to populate its metadata template. Such is the case for the geospatial datasets incorporated from the Geospatial One Stop (GOS) (geoData.gov). The Federal Geographic Data Committee (FGDC) Content Standard for Geospatial Metadata (FGDC-STD-001-1998) is required by and used to catalog the data in GOS. To enable incorporation into Data.gov of all data and tools published through GOS that meet the *Data.gov data policy*, an application has been developed that allows population of the Data.gov metadata template directly from the FGDC record that currently resides in GOS. GOS-published data exposed via the *Data.gov Geodata Catalog* thus feature both the standard Data.gov metadata, as well as the full FGDC content standard maintained in GOS. In the future, Data.gov and GOS will continue to integrate, with the ultimate goal of a merger between the two sites.

With regard to geospatial datasets, Data.gov also provides access to specialized data extraction tools (e.g., USGS Global Visualization Viewer) providing datasets in specific image data file formats such as satellite imagery and aerial photographs. These imagery data include additional metadata that describe relevant information about the data such as the spectral content, geospatial coordinates, image data quality (e.g., cloud cover), and other relevant attributes.

2.4. Growing the number of Datasets Published via Data.gov

A March 11, 2009, memorandum requested that Federal Chief Information Officers (CIOs) provide information on their Agency datasets that would potentially be suitable for the Data.gov initiative. This initial data call yielded 76 datasets and tools from 11 agencies that were made available on Data.gov when it was launched on May 21, 2009.

Since the initial launch, over 272,976 individual data resources have become accessible through Data.gov. These data resources include structured data and tools for the public to visualize and use data. Additional datasets continue to be added through the Data.gov dataset submission process. On July 3,

2009 Data.gov added a “Geospatial” catalog that houses both structured data and tools. Table 5 details the July 14, 2010, view of datasets, tools, and geospatial data on Data.gov from each participating Agency/organization. Note that some “tools” counted in Table 5 actually represent hundreds, thousands, or millions of datasets. For instance, the Department of the Interior’s US Geological Survey has a tool on Data.gov called “*USGS Global Visualization Viewer for Aerial and Satellite Data*”. This tool alone represents ten million archive images of the Earth's surface.

Our vision is to continuously improve and update Data.gov with a wide variety of available datasets and easy-to-use tools based on user feedback. The requested datasets that contain sensitive information (e.g., personally identifiable information, national security), are limited by technology (e.g., not machine readable), or that do not belong to the Executive Branch of the Federal Government are not available on Data.gov.

We received approximately 900 suggested datasets from the time of the site launch in May 2009 through December 2009. Representatives from the identified Federal agencies reviewed the suggested datasets from the public. Their responses fell into four categories: 16 percent of the data is already published on Data.gov (Already Published), 26 percent of the suggestions can be published in the near future (Actionable), 36 percent of the suggestions could be published at a later date (Potentially Actionable), and the remaining 22 percent of the suggestions cannot be published due primarily to security, privacy, or technology constraints (Not Actionable).

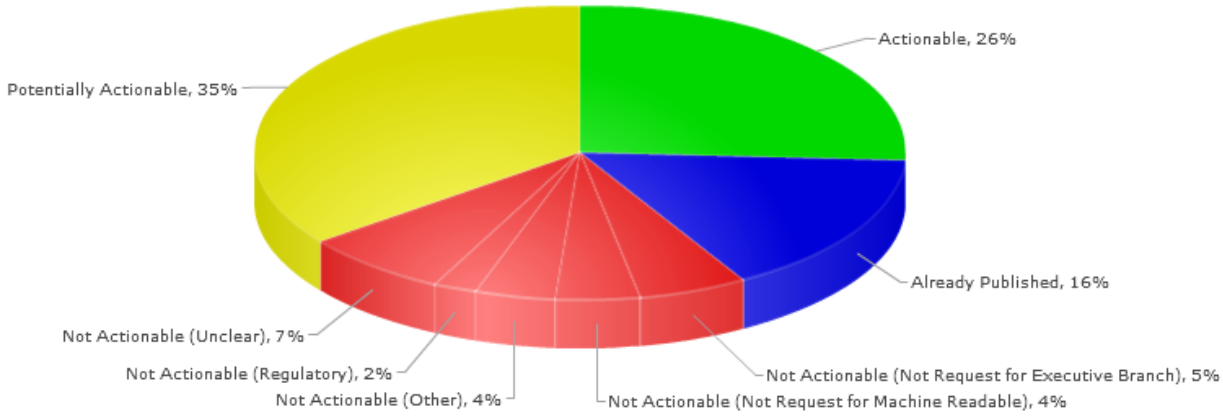

Figure 5: Agency Determinations of Suggested Datasets

Data.gov now provides these metrics on datasets, tools, and geospatial data at *http://www.data.gov/metric*. In interpreting Table 5 note that the *Open Government Directive* specifically required agencies to register at least three *new* high-value datasets on Data.gov by January 22, 2010. While many of the datasets submitted to Data.gov both before and after the January 22 deadline are high value, agencies have specifically identified the number of datasets in parentheses as "high-value" datasets in accordance with Open Government Directive provisions.

While Data.gov is proud of the quantity of datasets available, we continue to work with citizens and Agencies to identify and publish more and more data. Specific efforts along these lines are described in Section 3.

Table 5: Datasets and Tools by Agency/Organization (as of 7/14/10)

Agency	Raw Datasets (high-value)	Tools (high-value)	Geodata	Total
Department of Agriculture (USDA)	8 (3)	12	1	21
Department of Commerce (DOC)	65 (4)	92 (2)	166,494	166,651
Department of Defense (DOD)	18 (10)	196	0	214
Department of Education (ED)	3 (3)	16 (3)	0	19
Department of Energy (DOE)	57 (6)	16	0	73
Department of Health and Human Services (HHS)	76 (3)	64 (1)	0	140
Department of Homeland Security (DHS)	47 (3)	3 (1)	0	50
Department of Housing and Urban Development (HUD)	5 (5)	9	0	14
Department of the Interior (DOI)	190 (4)	7	103,558	103,755
Department of Justice (DOJ)	67 (3)	6	0	73
Department of Labor (DOL)	45 (6)	3	0	48
Department of State (STATE)	8 (3)	4	0	12
Department of Transportation (DOT)	3 (3)	10	0	13
Department of the Treasury (TREAS)	87 (3)	6	0	93
Department of Veterans Affairs (VA)	28 (3)	1	1	30
Environmental Protection Agency (EPA)	436 (4)	38 (4)	155	629
General Services Administration (GSA)	24 (7)	26 (16)	0	50
National Aeronautics and Space Administration (NASA)	3 (3)	21 (5)	515	539
National Science Foundation (NSF)	23 (2)	3	0	26
Nuclear Regulatory Commission (NRC)	4 (3)	0	0	4
Office of Personnel Management (OPM)	31 (2)	1	0	32
Small Business Administration (SBA)	3 (2)	2 (1)	0	5
Social Security Administration (SSA)	22 (14)	0	0	22
US Agency for International Development (USAID)	3 (3)	5	0	8

In order to facilitate Agencies' ability and willingness to enter their datasets into the metadata catalog, Data.gov is continually enhancing its capabilities to ease the burden on Agency POCs and provide them with multiple mechanisms to publish their data. When planned enhancements are in place, each Agency will have three mechanisms to publish metadata records to Data.gov (Figure 6). These three mechanisms are:

1. The Dataset Management System – this is a protected website only accessible by authorized users as described in Section 3.2.2. This website enables agencies to publish metadata records to the Data.gov catalog in accordance with the Agency's dissemination process.

2. A Publisher API – an application programming interface that will allow an Agency the ability to programmatically submit one or more records into the Data.gov catalog.
3. A Metadata Feed – if an Agency desires to control publishing of metadata records on their own websites, the Data.gov PMO harvesting service will read the metadata feed and publish the records to the Data.gov catalog. The metadata feed will be a file in a standard feed format like Really Simple Syndication (RSS) or the Atom Syndication format.

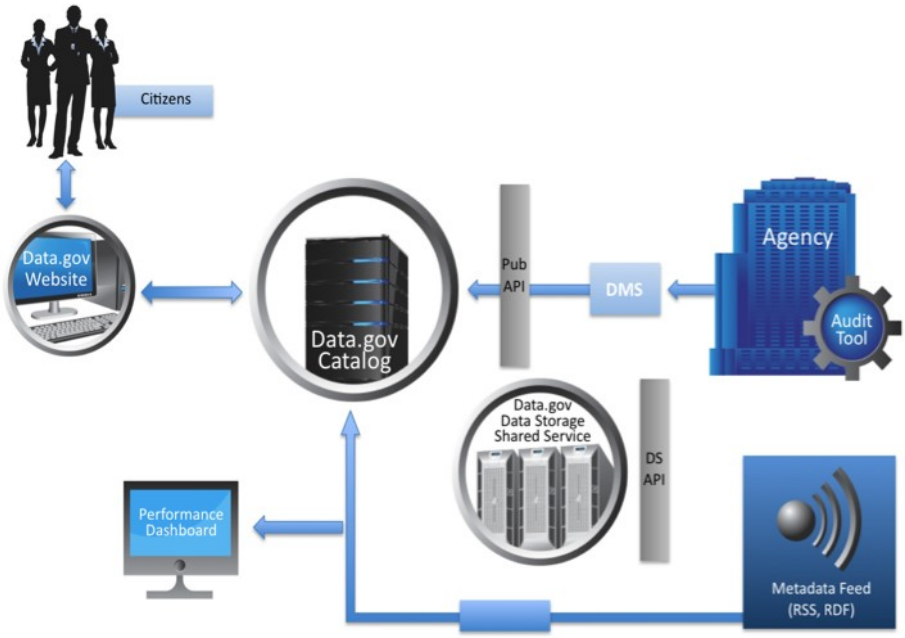

Figure 6: Publishing Architecture

Data.gov understands that the value the program provides depends on both the quality and the quantity of data available. Every effort is being made to streamline the publishing process and improve the tools that POCs can use to publish their Agencies' datasets to Data.gov.

2.5. Roles and Responsibilities for Data.gov Operations

Data.gov is an Executive Branch initiative that requires guidance, oversight, and expertise from a broad array of Federal organizations. The Data.gov governance framework outlines the associated governing bodies and their decision-making authority relative to the Data.gov initiative. The Data.gov governance framework is designed to provide guidance, oversight, and expertise for all functions related to the delivery of the Data.gov solution including management of business and information requirements, management of data standards, design and implementation of the Data.gov solution, and measurement and management of overall performance. The governance framework provides a formalized definition of decision-making roles and responsibilities for new and existing federal councils and committees, advisory boards throughout the process, and the roles and responsibilities of the Data.gov project manager and working group.

Individual teams identified within the governance framework are governed by their respective charters. This document merely provides overall context and guidance for delineating roles and responsibilities between the various governance bodies identified in this framework.

Table 6 summarizes the individual entities that comprise the overall Data.gov governance framework.

Table 6: Data.gov Governance Framework

Group / Role	Overview of Roles and Responsibilities
Executive Sponsor and Business Owner *Vivek Kundra* *Federal CIO*	Senior-most-executive-level official with decision making authority for the Data.gov initiative. The executive sponsor and business owner serves as the champion for Data.gov and is the final authority on the vision for the initiative and its intended results.
Data.gov Executive Steering Committee (DESC)	The DESC is designed to provide executive leadership, direction, and fundamental support for the project including review and acceptance of the architecture. The DESC makes policy and strategy recommendations to the executive sponsor and business owner. The DESC consists of the executive sponsor and business owner, the Data.gov Lead, volunteer representatives from the Federal CIO Council, the Manager of the Federal Enterprise Architecture Program Office, and executive from the Host Agency . The DESC works to promote collaboration and an exchange of ideas to build consensus for the vision and direction of Data.gov. In the event consensus cannot be reached, decisions will be made by voting. Each member of the DESC has one official vote.
Data.gov Initiative Lead	The lead for the initiative is a senior official with the authority to make decisions on the Data.gov initiative. The lead is responsible for interpreting the guidance, oversight, and expertise from the DESC to provide direction and operational controls to the Project Manager. The lead secures funding, appoints key personnel, approves the completed tasks, and makes other decisions as required.
Data.gov Project Manager (PM)	The PM is the operational leader for project activities and is responsible for achieving the overall Data.gov objectives. The PM is responsible for executing the vision defined by the DESC and articulated by the co-leads. The PM is responsible for organizing efficient and effective day-to-day management and operations of the Data.gov project.
Data.gov Project Delivery Teams	The Data.gov Project Delivery Teams take direction from the Data.gov PM to execute the project plan and activities and fulfill the Data.gov objectives. The Data.gov Project Delivery Teams are responsible for the full lifecycle of the project from business requirements through implementation.
Federal CIO Council	The Federal CIO Council provides executive level counsel to the Data.gov project via the Data.gov co-leads and project manager.
Data Architecture Sub-Committee (DAS)	The DAS provides the Data.gov project with data related guidance on a review basis. The Data.gov Project Delivery Teams will produce deliverables that are reviewed by many of the advisory teams including the DAS.
American Public	Data.gov thrives on citizen input. Public feed-back venues described in Section 1 provide mechanisms by which any member of the public can provide feedback and suggestions for the Data.gov initiative.

Agency Roles and Responsibilities

Agencies, via their data stewards, are the key partners responsible for populating Data.gov with high value, authoritative data. Departmental and Agency CIOs were asked to appoint designees, referred to here after as " points of contact" (POC) to coordinate both within their Agency and between their Agency and the Data.gov Program Management Office (PMO) run by GSA. Within Departments, many individual operating units (Agencies, Offices, and Bureaus) identified their own POCs. The POCs have direct support from the Data.gov PMO office as appropriate so the POCs can focus on serving their own data stewards. The POCs are key points of coordination as they help form a bridge between those Agency staff that steward the datasets and the Data.gov PMO. The term "data steward" is used to refer to the Agency staff that is directly responsible for managing a particular dataset. The Agency's data stewards are the source of the documentation for each dataset as well as assurances regarding the quality of that data.

Agencies' functional roles related to Data.gov are described below:

- Agency administrators, in support of enterprise transparency, should direct all parts of their organization to jointly coordinate and support Data.gov requirements as part of their enterprise architecture efforts.
- Agencies are encouraged to vet Data.gov requirements internally. To accomplish this goal, agencies may wish to convene a Data Steward's Advisory Group whose participants would represent each of the Agency's key mission, business, and organizational areas. This would have the effect of empowering individuals Agency-wide who are most familiar with potential datasets that could be made ready for public dissemination.
- Agencies are the source of the data that are posted to Data.gov.
    - Agencies are responsible for determining which data and tools are suitable to be posted on Data.gov. As they select datasets for Data.gov they should be mindful of the significance of the increased profile the data will have as part of a high profile Presidential Initiative. Agencies are encouraged to promote authoritative data sets and document the quality of those datasets.
    - Agencies retain the right and responsibility for managing their own data and providing adequate technical documentation. This role tends to be carried out by program offices within the context of their particular missions. The term "data steward" is used to refer to the Agency staff that is directly responsible for managing a particular dataset.
    - Agencies, in conjunction with Data.gov POCs and CIOs, are responsible for ensuring that their data assets are consistent with their statutory responsibilities within the context of information dissemination, including those related to information quality, national and homeland security, accessibility, privacy, and confidentiality.
- Agencies are responsible for ensuring that the data stewards for a particular data asset complete the required metadata for each dataset or tool to be publicized via Data.gov.
- Agencies should facilitate Data.gov POCs and CIO efforts to understand and catalog data assets, as indicated below.

- Agencies are responsible for cataloging[18] and understanding their data assets, establishing authoritative sources, and ensuring the high quality of data. Agencies are encouraged to engage their enterprise architecture programs to formally catalog their data assets, determine which sources are authoritative, and evaluate adherence to information quality guidelines. Agencies are encouraged to leverage the *Federal Data Reference Model* which provides agencies with assistance to:
    - Identify how information and data are created, maintained, accessed, and used;
    - Define Agency data and describe relationships among mission and program performance and information resources to improve the efficiency of mission performance.
    - Define data and describe relationships among data elements used in the Agency's information systems and related information systems of other agencies, state and local governments, and the private sector[19].
- Agencies have the responsibility of ensuring that authoritative data sources are made available in formats that are platform independent and machine-readable. Agency enterprise architecture programs should promote the publication of web services, linked open data, and general machine-readable formats such as XML.

Data.gov has established a network of over 250 Agency and bureau-based POCs. POCs were polled to share their key success factors. The following were provided by POCs from EPA, Commerce, Interior, Defense and Justice:

- Full backing of and established working relationships with Agency senior executives.
- Robust processes and governance that recognize specific mission(s) and reporting structures.
- Proactive data stewards with thorough knowledge of the submission process, quality requirements, and online tools.
- Agency designated leads with breadth of experience and solid knowledge of the Agency mission and its data holdings.
- The ability to understand and explain the benefits of open government concepts and information sharing in terms of the Agency's mission and support organizations.
- An active Data.gov POC Community of Interest (COI) to collaborate, share lessons learned, and best practices.

[18] Establishing, maintaining, and publishing inventories, priorities, and schedules of all Agency dissemination products is a requirement established by *OMB Memorandum M-05-04*, "Policies for Federal Agency Public Websites." *OMB Memorandum M-06-02* further clarifies this by stating that "in addition to publishing inventories of specific information dissemination products, agencies must also publish inventories of other information to which public access is appropriate. In deciding what other information to include on an inventory and permit access by the public, agencies should take the broadest possible view and assume at least some members of the public or specific users will be interested in the data. Such additional information could include databases of underlying data even though actual use would require a high degree of sophistication. Again, it could be appropriate for an Agency to make this information available, but let the market determine what value added services are desired by the public."

[19] OMB Circular A-130, "Management of Federal Information Resources," section 8b2(b)(iv).

- Agencies have the responsibility for assigning an overall Data.gov "point of contact" (POC) for their Agency. The Agency's Data.gov POC is responsible for ensuring that the requested documentation accompanies all datasets posted to Data.gov.
- The POC is responsible for training data stewards as to the importance of the metadata template that accompanies a Data.gov submission as well as how to complete this template.
- The data steward has the responsibility for:
    - Documenting the Agency's data using the Data.gov metadata template.
    - Ensuring program sign-off has been obtained for submission of the dataset, and working with the POC, if necessary, to obtain Agency sign-off for posting on Data.gov.
    - Ensuring that the data are available online through the Agency's website.
- The POC is responsible for understanding Data.gov processes, Data.gov metadata requirements, and compliance requirements for coordinating data submissions for the Agency. The Data.gov POC role is expected to evolve as each Agency's dissemination processes mature and the Data Management System improves (as discussed in the next section). Initially, the role and responsibilities of these POCs are as follows:
    - POCs are responsible for coordinating an internal (Agency) process that identifies and evaluates data for inclusion in Data.gov. Such a process must include: a) screening for security, privacy, accessibility, confidentiality, and other risks and sensitivities; b) adherence to the Agency's Information Quality Guidelines; c) appropriate certification and accreditation (C&A); and d) signoff by the program office responsible for the data.
    - The POC should help coordinate the exposure of this metadata to Data.gov in one of the several ways detailed in this Concept of Operations document.
    - The POC is also responsible for facilitating feedback to Data.gov from Agency personnel regarding improvements to the metadata requirements, including recommendations that generate taxonomies to facilitate interoperability.
- Critical success factors for Agency participation in Data.gov include:
    - Establishing, populating, and moving Agency data and tools through a "pipeline" culminating in inclusion in Data.gov.
    - As Data.gov matures, utilize feedback on the value generated and feedback garnered via Data.gov to improve Agency participation.
    - POCs' success in serving as a conduit for Agency participation in the evolution and successful realization of target outcomes for Data.gov.

OMB's Senior Advisory Group

OMB's development of Data.gov data policies is informed by advice senior-level government employees who primarily represent formally chartered inter-Agency working groups that focus on various aspects of data policy, encourages the development and implementation of a unified vision for achieving data interoperability and other efforts to modernize Executive Branch data dissemination and sharing. The Executive Branch data policy experts are referred to as OMB's Senior Advisory Group (the Advisory Group). The Advisory Group provides OMB with a forum for working interactively with senior program executives, Agency data stewards, and others responsible for the generation and dissemination of data accessible through Data.gov.

The Advisory Group currently has representatives of the following communities:

- The *Chief Information Officers' (CIO) Council*, which includes the CIOs from all Chief Financial Officer-Act departments and agencies.
- The Inter-Agency Council on Statistical Policy (*ICSP*), which includes representation from 14 principal Federal statistical agencies;
- The Federal Geographic Data Committee (*FGDC*), an inter-Agency committee that promotes the coordinated development, use, sharing, and dissemination of geospatial data on a national basis;
- The Commerce, Energy, NASA, and Defense Information Managers Group (*CENDI*)**,** an inter-Agency working group of senior scientific and technical information (STI) managers from *13 Federal agencies;*
- The Inter-Agency Working Group on Digital Data (*IWGDD*), coordinated by the Office of Science and Technology Policy (OSTP);
- The Networking and Information Technology Research and Development (*NITRD*) Program, the Nation's primary source of federally funded revolutionary breakthroughs in advanced information technologies such as computing, networking, and software.
- The U.S. Group on Earth Observations (*US GEO*) is driving the interoperability of land-, sea- air- and space-based Earth observations across 17 federal agencies.

Specifics on the Advisory Group include:

- The Advisory Group is designed to be a vehicle for cross-government fertilization with respect to information policy and transparency opportunities raised by the Data.gov initiative as well as a mechanism for mobilizing support for and implementing the transformational goals of Data.gov. Furthermore, the Senior Advisory Group helps the Data.gov team understand and establish models, frameworks, and technology support for performance measurement and management around information quality, dissemination, and transparency.
- The Advisory Group is one of several vehicles that the Office of Management and Budget (OMB) uses to obtain advice on the direction of Data.gov. OMB also receives advice from the Project Management Office at GSA, the CIO-appointed Agency POCs, representatives of related websites described in Section 4, as well as from technology and information policy leaders outside the government, and Data.gov users.
- In addition to representing the perspective of their Federal community of interest, Advisory Group participants will be asked to speak from the perspective of Agency data stewards, providing feedback to OMB on the potential impact of Data.gov proposals on Agency data generation and dissemination programs.
- The Advisory Group is not a decision making or voting body, and consensus will not be sought.
- The Advisory Group is co-led by OMB's Office of Information and Regulatory Affairs and OMB's Office of E-government and Information Technology.

Data.gov Program Management Office (PMO)

The Data.gov PMO provides effective project management services to Data.gov. These include monitoring cost, schedule, and budget; preparing required management reports to GSA and OMB

leadership (e.g. OMB exhibit 300 and segment architecture blueprints); requirements management using the JIRA tool in OMB's MAX collaboration website; coordination among the various Modules; and interaction with the Data.gov Executive Steering Committee.

Several IdeaScale suggestions were relevant to the operations of the PMO:

- One user said that it is essential to have "coordinated change management… with change champions in all impacted agencies governed by a central program management office for messaging, quality, and approach" so that Data.gov meets its goals. Agencies can be reluctant to change their data management methods without some leadership. Effective project management and leadership for Data.gov can help champion this change. In fact, the Data.gov PMO is actively coordinating with Agencies through the Agency POCs to help bring consistency and interoperability to Agencies' data management policies.
- There was a caution from one participant about making Data.gov and the Open Government Directive into "unfunded mandates." This user said that "The requirement to make data accessible to (through) Data.gov should be formally established as a component of one of the Federal strategic planning and performance management frameworks… and each Agency should be funded (resourced) to help ensure Agency commitment towards the Data.gov effort. Without direct linkage to a planning framework and allocation of dedicated resources, success of Data.gov will vary considerably across the Federal government." If the strategic planning framework is a route Data.gov chooses to take, effective project management will be needed.

Subject Matter Expert Technical Working Groups

Data.gov will further harness the interests and expertise of staff across the government when it stands up technical working groups designed to develop approaches needed to further specific goals of Data.gov. This includes modernizing and streamlining data formats and structures to allow linking, tagging, and crawling. For instance, the Data.gov team will draw on expertise from across the government to provide advice regarding the best approaches to publishing metadata that facilitate encoding meaning into datasets in such a way that they are directly interpretable by computers and strengthen the interoperability of Federal datasets.

Federal Communities of Interest/Information Portals

Other inter-Agency efforts, such as Science.gov and Fedstats.gov, are focused on serving distinct user communities and to make information easier to find and more useful for those communities. These Federal communities of interest often disseminate their data through information portals. Increasingly, these sites will have the opportunity to mimic the design patterns of Data.gov including the metadata template, catalog capabilities, and end user search and feedback capabilities.

Federal communities of interest that offer these information portals to the public are encouraged to first standardize a metadata taxonomy or syntax to be shared with Data.gov, and then communicate any changes to it as the community evolves the standard. These communities of interest are encouraged to expose corresponding data either as downloadable data, query points, or tools. In this way, the information portals provided by these Federal communities of interest will become more standardized in how their data is maintained and shared, and these information portals will become networked to Data.gov to allow for maximum visibility, discoverability, understanding and usefulness of the data.

3. Data.gov's Collaboration-Driven Conceptual Solution Architecture

The current state physical architecture for Data.gov consists of a website and a relational database that serves as the metadata catalog containing the site's content. The evolution of the physical architecture will be based on the conceptual architecture that is depicted in this document.

The public feedback on the current Data.gov architecture has been fairly uniform in that:

- Federal agencies that produce data want an easier way to make their data available on Data.gov
- End users of Data.gov want easier ways to use the metadata from Data.gov and the actual Agency data represented on Data.gov.

> The conceptual architecture has been developed based on feedback from the user community, feedback from the data producing Federal agencies, and overall alignment with Data.gov's strategic intent and core design principles outlined in Section 1 of this document.

3.1. Evolution of the Data.gov Website

The Data.gov website has undergone substantial growth and maturation in its first year of operation. In order to provide the best possible value to the citizens and the government, Data.gov will continue to evolve. This section presents a brief history of the website and the changes it has undergone since launch.

3.1.1. Original Site

The Data.gov website was launched on May 21, 2009. The original website presented users with a clean, simple, and easy to use interface (Figure 7).

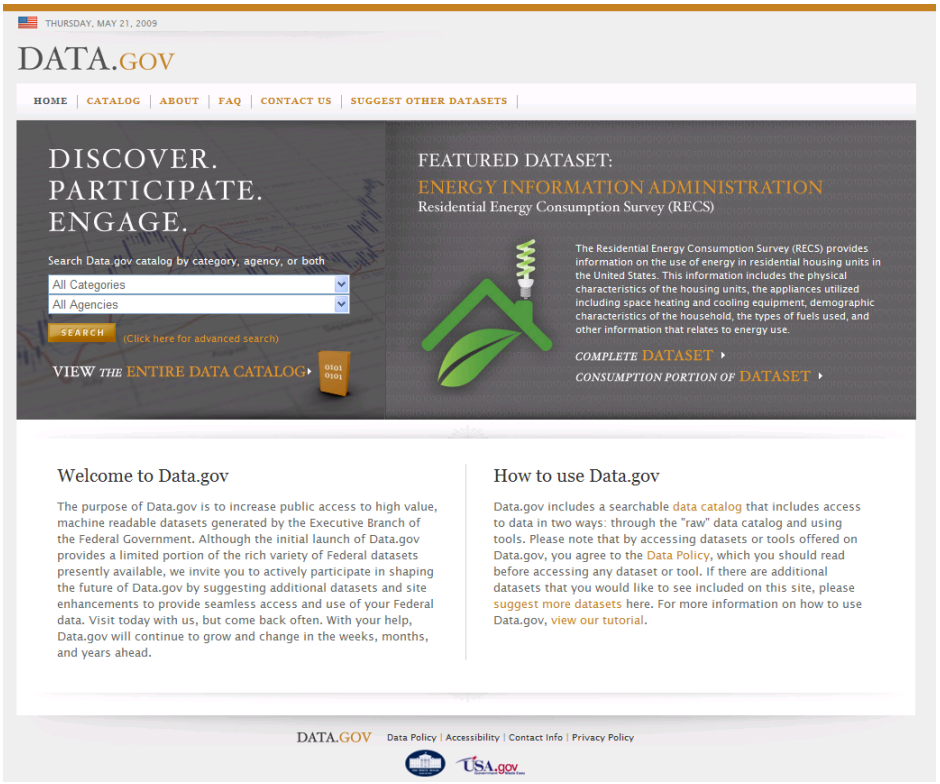

Figure 7: Original Data.gov Homepage

Within five hours of launch, there were over 1000 tweets and over 300 suggestions from citizens about the site. Response was largely positive, with comments such as:

- "Yes, we can have open data!"
- "Data.gov is now live! It's got more of a Web 2.0 feel than a government bureaucracy."
- "Nice! Data.gov just launched, accumulating government data for everyone to explore"
- "US Government data out the wazoo! http://www.data.gov/ launches with a snazzy design (for once). Looks like a good tool to me"

Data.gov started with 47 datasets and 27 tools. On its first day, Data.gov received 2.1 million hits; on the second day, 2.5 million. Within the first two months, it had received over 20 million hits from all over the world.

3.1.2. July 2009 Release

Despite its initial success, Data.gov did not rest on its laurels. User suggestions for additional machine readable formats, web services/API to directly access and use the data, and more datasets did not go unheeded. Neither did requests and suggestions for new functionality and features. The original site provided citizens access to a catalog of government data, which they could browse by category (aligned to the categories of the Statistical Abstract of the United States) or by publishing organization. The search function of Data.gov was primitive, and no distinction was made among the "raw" data, tools, or geodata catalogs. The site evolved rapidly, and by July 2009, a new version of the website came online (Figure 8).

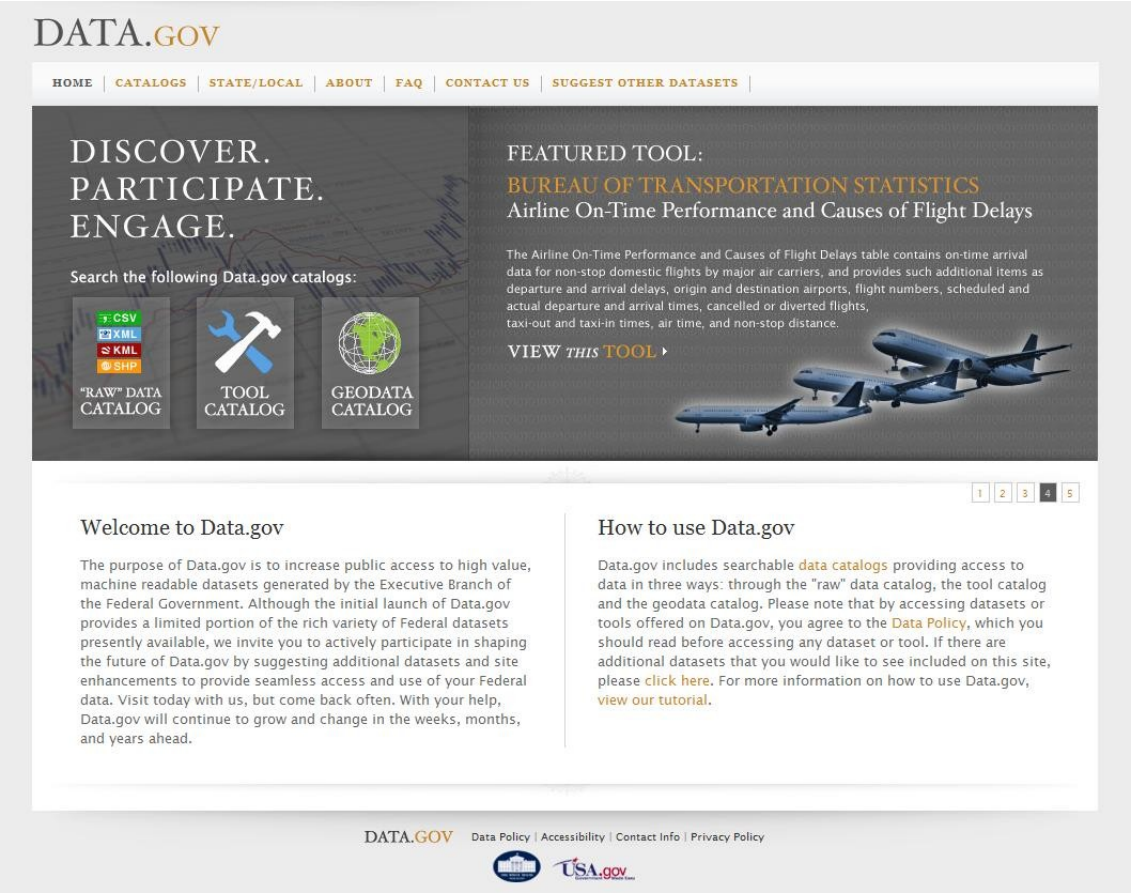

Figure 8: July 2009 Data.gov Home Page

While cosmetically, the new website looked similar, enhancements had been made to the features and functionality of the site. The datasets were now organized into the three catalogs (raw data, tools, and geodata), and links to state and local government data sites were also provided. The new homepage also featured rotating panes of new and interesting datasets and tools.

Inside the homepage, additional functionality enabled users to filter datasets by category and Agency, and a five-star dataset rating system was introduced, enabling users to rate each dataset on utility, usefulness, ease of access, and overall quality (Figure 9).

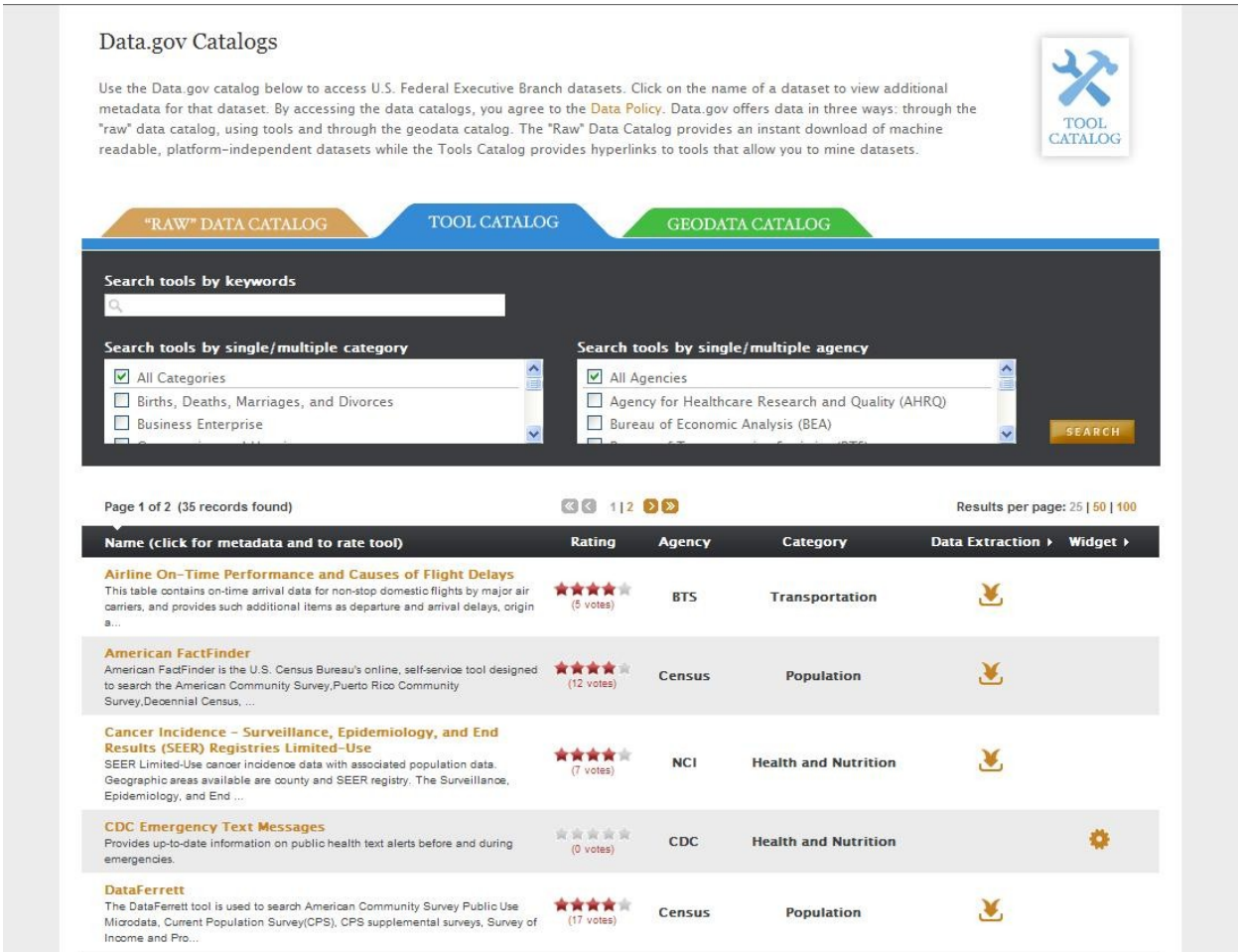

Figure 9: July 2009 Data.gov Catalog Page

User feedback continued to drive many of the improvements to Data.gov. For example, one user suggested that "The system should have a way to add 'notes' about the data." This suggestion was incorporated and the new release provided citizens with a mechanism to comment on each dataset or tool on the metadata page (Figure 10).

The metadata page provides descriptive information about each dataset or tool. It is completed by the submitting Agency POC, and enables users to see what is in the dataset, where it came from, when it was published, and other pertinent facts.

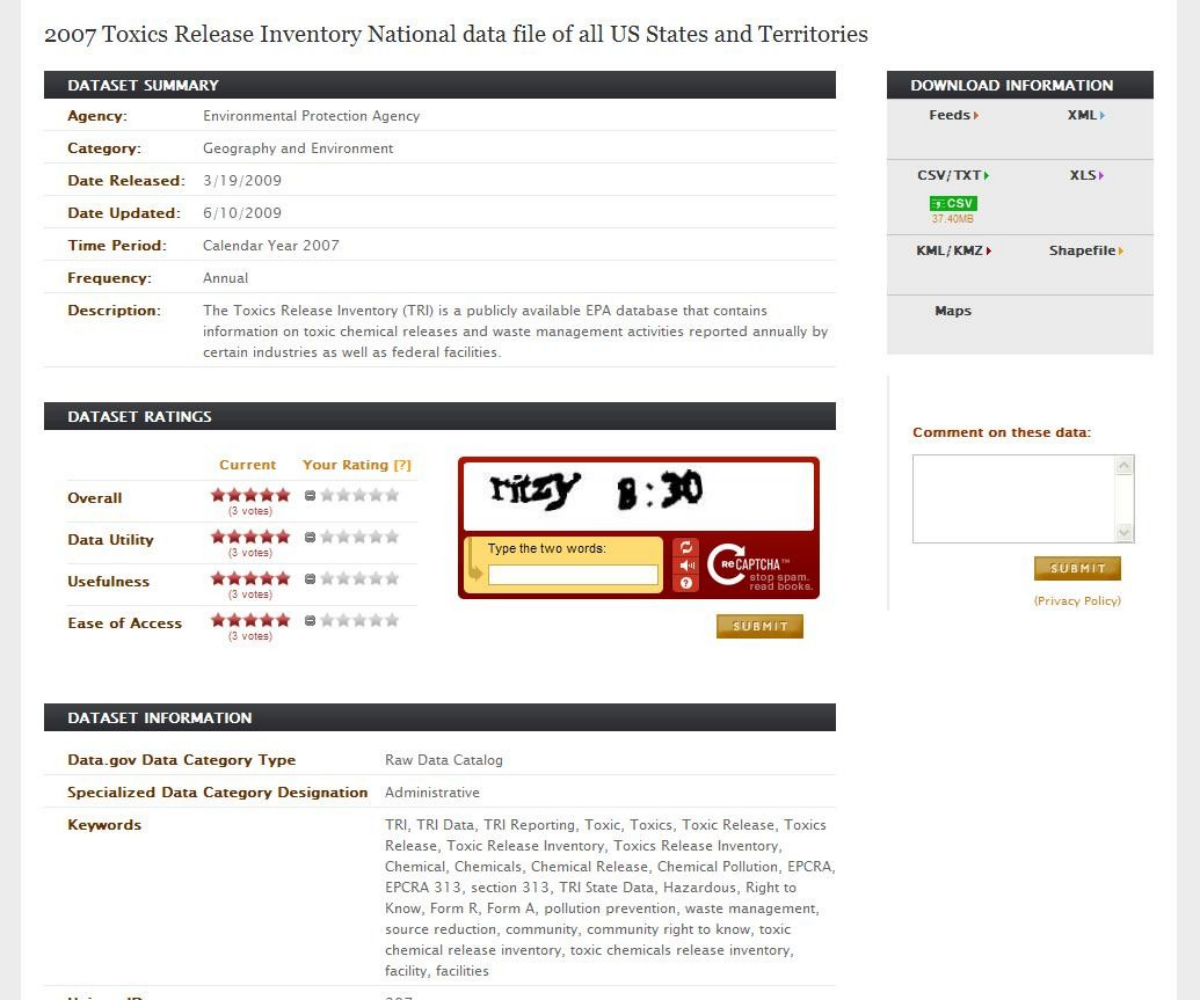

Figure 10: July 2009 Data.gov Metadata Page

3.1.3. March 2010 Release

Data.gov did not stop there. A new version was released in March 2010 that provided a significant redesign of the Homepage, as well as still more enhancements to the functionality, most notably metrics reports for the site, and integration of the search function with Search.USA.gov, greatly improving users' ability to find datasets using keyword searches (Figure 11).

Figure 11: March 2010 Data.gov Home Page

The metrics available on the new Data.gov fell into three categories: Federal Agency, Download Statistics, and Visitor Statistics. The Federal Agency statistics page provided the number of raw datasets, tools, and geodata datasets by Agency, as well as the date of the latest upload and the number of times each Agency's datasets were accessed within the last week. The Download Statistics page showed the number of times each category of dataset had been downloaded. Finally, the Visitor Statistics page provided a number of reports on the number and location of the site's customers. Sample metrics reports are shown in Figure 12.

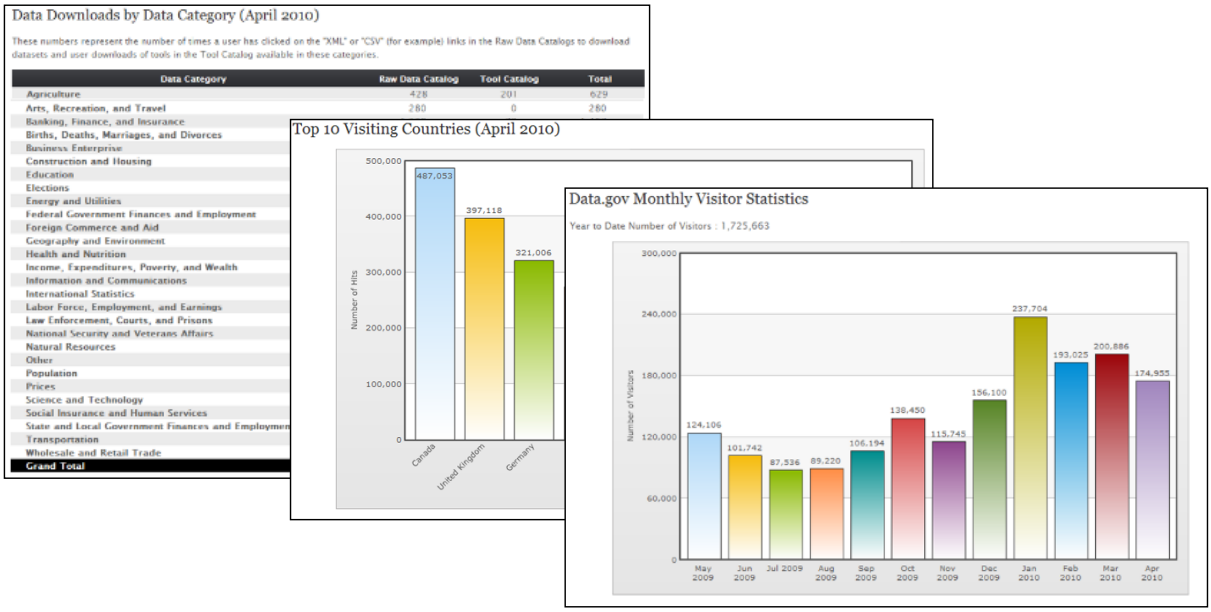

Figure 12: March 2010 Data.gov Sample Metrics Reports

One citizen had suggested "The current Data.gov search tool is capable but cumbersome and does not yield desired results. We should combine and leverage capabilities in other government websites that have good search engines (and/or are planning on more improvements such as usa.gov)."

USA.gov is an easy-to-search, free-access website designed as a centralized place to find information from U.S. local, state, and federal government Agency websites. USA.gov's objective is to provide a free service, enabling the global community to easily and rapidly find U.S. government information that has been posted on the Internet. It offers a powerful search engine and an index of web-accessible government information and services to help citizens find the information they need. Incorporating Search.USA.gov into the Data.gov website has been one of the most successful and popular upgrades in Data.gov's brief history.

The March release also provided citizens the ability to look specifically at the "high value" datasets provided by each Agency required by the Open Government Directive. By highlighting these datasets, Data.gov enabled users to assess for themselves each Agency's compliance with the Directive.

This version of the site introduced the Data.gov Blog, a means for Data.gov leaders to communicate directly to citizens; the Developers' Corner, a forum for citizens and developers to share their ideas and the apps, mash-ups, and widgets they developed to put Data.gov's data to use.

3.1.4. First Anniversary Release

The March Data.gov release added features, but user feedback indicated that it was perhaps overly cluttered and confusing to new users. In conjunction with the first anniversary of the launch of the original Data.gov, a new version was released with a streamlined look and feel, and a new set of features, again responding to user suggestions received through IdeaScale (Figure 13).

Figure 13: May 2010 Data.gov Home Page

By the first anniversary, Data.gov had grown to over 270,000 datasets (from 47 at launch), and was averaging over 200,000 visitors per month. In May, the month of the anniversary, there were nearly 21 million hits. The total number of hits to the Data.gov website since launch was over 100 million.

Among the new features of the latest version of the site are the Community page, which provides links to state, local, tribal, and international data sites, as well as related programs and organizations such as Open.gov, Rensselaer Polytechnic Institute (RPI), the Sunlight Foundation, and the World Bank; enhanced metrics, including a display of the disposition of suggested datasets; a current list of the most popular datasets; an "Apps Showcase" that displays some of the user-created applications that use Data.gov data; and Semantic Web resources such as a collection of Resource Description Framework (RDF) "triples" that support the developing "web of linked data." See more on Semantic Web in Section 3.3.1.

As with the previous updates and enhancements to Data.gov, many of the changes were in response to user comments and suggestions. For example, many users are interested in finding or publishing applications that take advantage of Data.gov. To quote one, “Create a page that highlights applications that are making use of data from data.gov. The examples should be real world (not demo's) sites that currently are making use of the data. They should range from simple to more complex. The public should be able to submit their sites to list.” Data.gov’s new Apps Showcase (Figure 14). Today’s Apps Showcase is only the beginning. Data.gov will build upon the success of the first year of making government data available, and emphasize making data useful. Apps, mash-ups, and widgets are a prime example of how citizens can benefit from the data published on Data.gov.

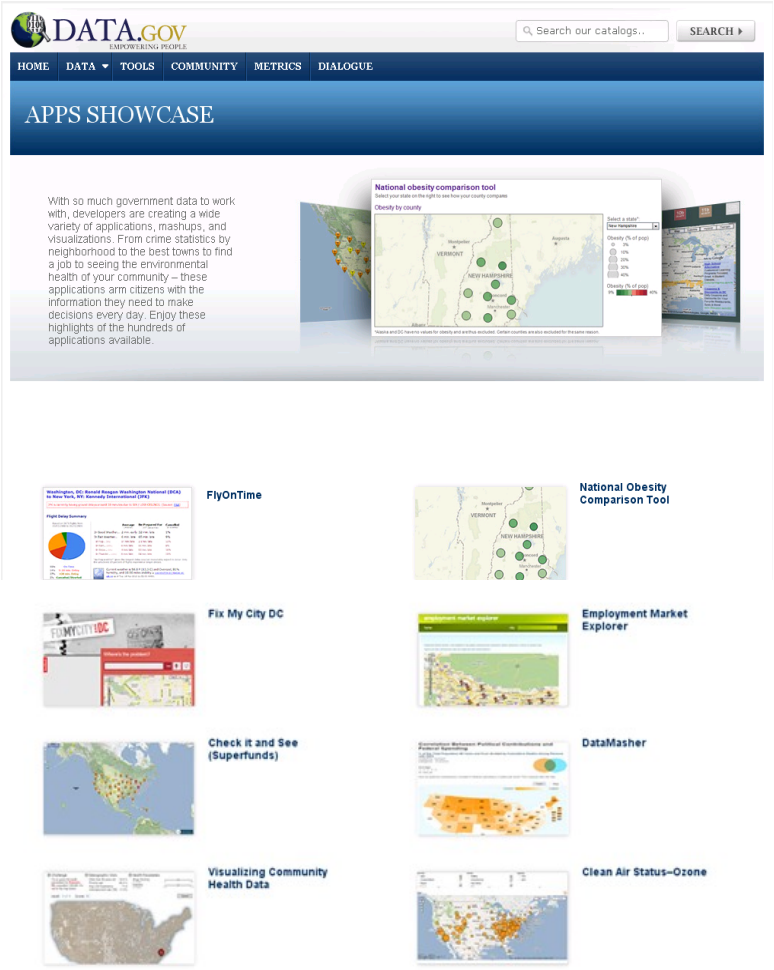

Figure 14: Data.gov Apps Showcase

An example of an app that uses data from Data.gov is FlyOnTime, which enables travelers to “find the most on-time flight between two airports or check how late your flight is on average, in good weather and bad, before you leave.” A mash-up combines the data from two different sources to provide useful new information. For example, a mash-up between education expenditures and unemployment, by state, might provide insight to the relationship between the two statistics. A widget is a mini-app that can be installed on a desktop or web page that automatically retrieves information such as weather alerts, product recalls, or other news. Data.gov supports and encourages the development of apps, mash-ups, and widgets that make data more useful to citizens.

Many of the new features of the anniversary release are available by clicking on dedicated buttons on the home page, but they can also be reached through the improved rotating highlight panes feature. The panes call visitors' attention to significant developments at Data.gov. The panes are updated regularly, and old panes are available in a gallery, shown in Figure 15.

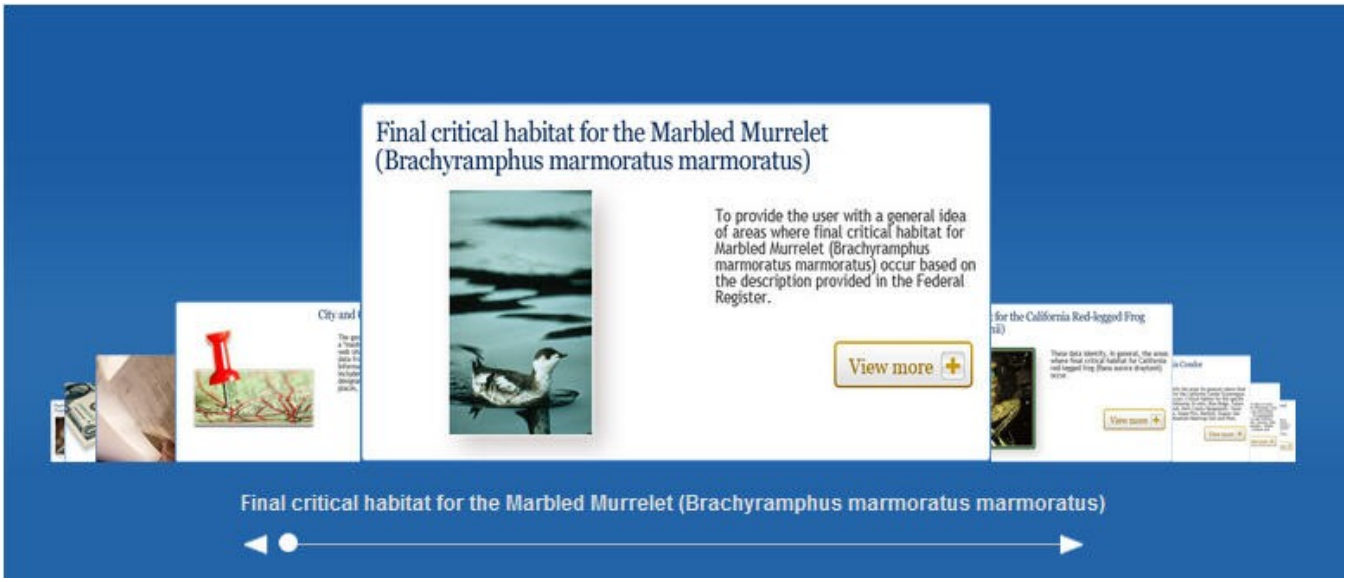

Figure 15: Data.gov Highlight Panes

Data.gov has matured significantly in its first year of operation, and collaboration with the user community has been a primary driver to many of the changes to the site and its capabilities. In the coming years, Data.gov will continue to grow, adding datasets to become the best source for government-produced data available. The features and functionality of the site and its infrastructure will continue to improve and remain on the leading edge of technology to provide citizens with not only access to data, but the tools to make that data useful to them.

As described in the following section, Data.gov's open, cloud-based architecture lends itself to continuous improvement in both capacity and capability. Developments in data hosting, semantic web, data discovery, and geodata visualization, among others, will enable Data.gov to realize the vision of open and transparent government by "democratizing" data; making it more available in "Internet time" from anywhere, at any time, readily and reliably.

3.2. Data.gov Technical Architecture

Data.gov evolved from initial concept to go-live in an intense period of visioning, conceptual design, analysis of alternatives, and decisions beginning in March 2009. The project team was stood up, strategy and policy written, budgets developed, and the technical infrastructure implemented in less than eight weeks. The rapid deployment of Data.gov clearly demonstrates the government's ability to be agile and innovative in an environment where technologies are often obsolete by the time they ever see the light of day.

The Data.gov team developed several alternatives for the conceptual architecture and evaluated them based on cost, speed to deploy, potential for future growth, data integrity, technical feasibility, and other factors (see Figure 16). Based on these criteria, the Distributed Architecture option was selected, and development of the website and its supporting infrastructure commenced.

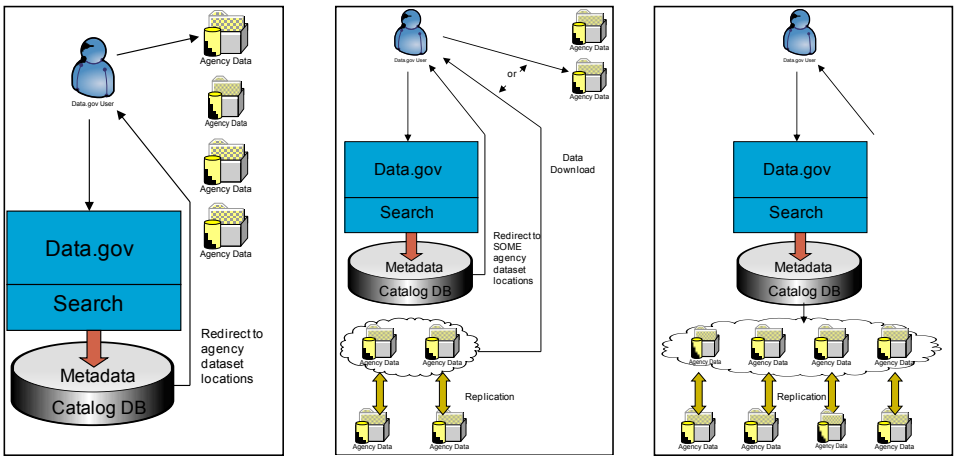

Figure 16: Data.gov Hosting Architecutre Alternatives

As depicted in Figure 17, Data.gov's technical architecture consists of six core modules: (1) the website; (2) the Dataset Management System (DMS); (3) the metadata catalog; (4) a performance tracking and analysis engine; (5) data discovery; and (6) a hosting service.

The modules will be made more accessible through a collection of application programming interfaces (APIs) that expose metadata and data. Together, these modules, tools, and APIs will allow Data.gov to adapt to its customer base as needed. Note that many of the capabilities, such as the Dataset Management System, are currently in use. Where this is the case they will be enhanced and extended. In other cases, for example the data infrastructure tools, the Data.gov team will partner with others to deliver the capability.

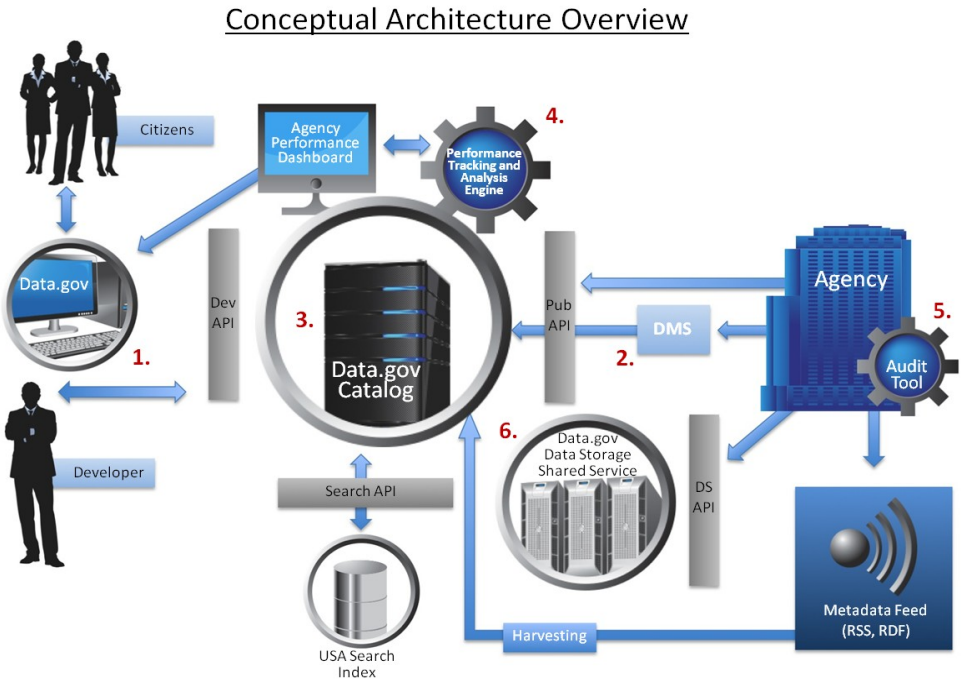

Figure 17: Conceptual Architecture Overview Diagram

3.2.1. How Does the Public Access Data.gov? Module 1 – Website and Search

The Data.gov website is publically available; anyone can discover data published by the federal government and download the data to their local computer. To serve up these datasets, the Data.gov website accesses a catalog of records with one record representing each dataset published to it.

The Data.gov website has a three-tiered design. The first tier features a home page that offers navigation to the catalogs and tools, community sites, metrics, and forums for public feedback. The choice of featured datasets and tools is replaced periodically, illustrating the datasets that support different missions throughout the government.

The second tier of the Data.gov website currently incorporates three catalogs:

1. "Raw" Data Catalog[20]: Features instant view/download of platform-independent, machine-readable data in a variety of formats.
2. Tool Catalog: Provides the public with simple, application-driven access to Federal data with hyperlinks. This catalog features widgets and data-mining and extraction tools, applications, and other services.
3. Geodata Catalog: Includes trusted, authoritative, Federal geospatial data. This catalog includes datasets and tools. This catalog employs a separate search mechanism.

Note that there is overlap among these three catalogs. For instance, there are some geospatial data in the “raw” data catalog and the Geodata Catalog technically contains both “raw” data and tools.

The third tier of the Data.gov website displays the information necessary for the data user to determine the fitness of the dataset for a given use. The source of that information is the metadata template completed for each dataset by the contributing Agency. Note that the end user sees the core metadata and the metadata associated with the domain specific metadata standards (e.g. geospatial, statistics) where those exist. This third tier view and features are shown in Figure 18. The end user accesses this page for links to download the data/tools and descriptions, as well as links to the source Agency. Also on this page, the end user can download the data and/or tools, rate the data, and even comment on the data or tool.

[20] The term “raw data” is used within the Data.gov context to mean data that are in a format that allows manipulation and are disaggregated to the lowest level consistent with maintaining privacy, confidentiality, and national security.

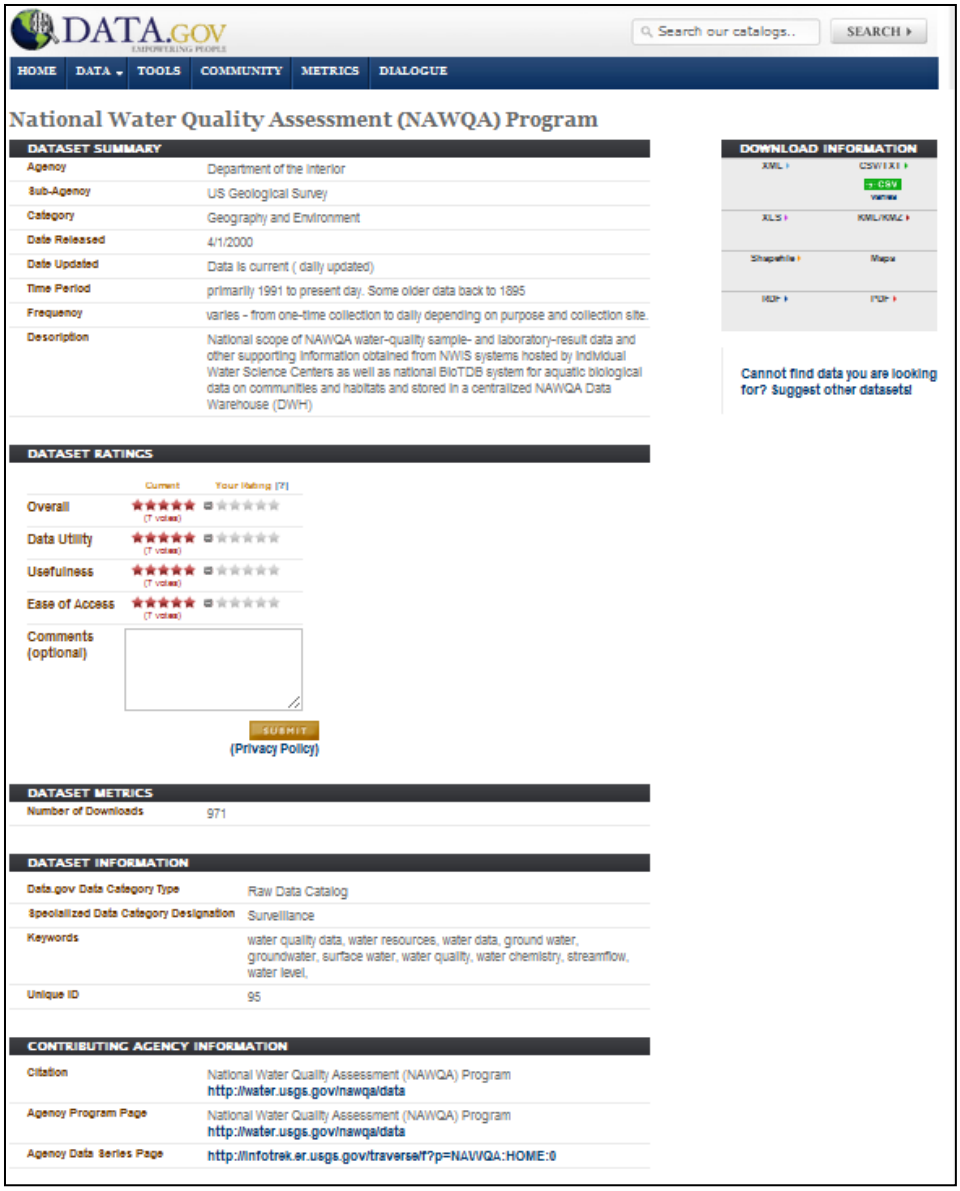

Figure 18: Data.gov Metadata Page

Many of the ideas submitted by citizens to IdeaScale were centered around improving the customer experience through “look and feel” enhancements.

- A user suggested that “An API to provide customizable RSS feeds should be considered, to allow users to subscribe to specific thematic areas, geographic areas and so on” so they can be notified by email or by RSS when new or updated data is posted. RSS feeds and email alerts to keep users informed of new or updated datasets are being planned for a future release. . A variation of this would be alerts to developers related to changes or updates to datasets they use to power their applications. Alerting and notification as a feature could be implemented via a data infrastructure tool, or via specific features added into core modules, or both. This is an area where Data.gov might implement a basic capability and invite experimentation and innovation to identify opportunities for greater added value – data domain specific, in general, or in some unforeseen manner.

- Many of the datasets on Data.gov are related to other datasets, by either time, geography, or some other connection. For example, the Toxics Release Inventory (TRI) data is provided in single files by year and by state. A participant suggested that it would be useful to link these sets and be able to treat them as a group. This idea is also connected to ensuring more consistent metadata amongst datasets and semantically linking information. As Data.gov implements the data hosting and semantic web enhancements, linking related datasets will become the norm rather than the exception.
- Participants were interested in more options and information to filter datasets and increased use of taxonomies to be able to have improved browsing for information. For example, a participant said, "There should be a way to filter by only data that is actively being updated or dynamic." Work is currently underway to establish more robust and flexible taxonomies and metatagged "folksonomies" that citizens can use to filter and find relevant datasets.
- Users contributed a few straightforward requests for improving Data.gov usability, including a call to making the download button on each data page more obvious, reducing the number of graphics, and simplifying language instructions throughout the site to be more straightforward. These suggestions have been incorporated and the Data.gov website continues to evolve to become more user-friendly and intuitive.

Data.gov is also improving search capabilities across the three Data.gov catalogs to allow the public to more easily find datasets. A few of the ideas and comments in the IdeaScale site are related to improving search within Data.gov's site by improving the interface and the metadata associated with datasets:

- One participant expressed frustration with the current Data.gov search, saying, "The current Data.gov search tool is capable but cumbersome and does not yield desired results." This participant suggested using the capabilities of other government websites with good search engines (e.g., USA.gov). A commenter said that if USA.gov is used, the "advanced search feature should be enhanced to enable selective searching" for specific file types (e.g., .xml, .xsd). In fact, Data.gov is now partnered with Search.USA.gov, which provides significantly improved search results of the Data.gov catalog.
- A commenter noted, "A simple first step [for enhancing Data.gov search] would be to expose an OpenSearch description interface… On Data.gov itself, the ability to sub-filter and build up more complex queries would be great, with a link to share." For more information on implementing this idea, see http://www.opensearch.org, and Project Blacklight for an example of faceting and building query user interfaces: http://projectblacklight.org.
- Another participant expressed interest in enhancing the taxonomy and ontology attributes of Data.gov so that data can be more easily located by the search engine: "One thing that might provide value, particularly to, for example, the science community might be integration and use of existing taxonomies, ontologies, and thesauri as developed by various agencies… this type of thing can be integrated into a search box via AJAX to facilitate more robust searches, or to dynamically broaden searches and navigate for greater or lesser granularity." This sentiment was echoed in many other conversations on the IdeaScale site, particularly as related to browsing and enhancing the metadata system. In the near term, Data.gov is developing taxonomies (a hierarchical structure of topics) that allow users to drill-down by topic area, improving a user's overall search experience. Data.gov's search capability will be improved by adding an advanced search feature, end-user tagging (known as a folksonomy) of datasets and the ability to "search inside" datasets for keywords. The advanced search feature will expand

the number of data types that can be selected for search to include XML, RDF and all other formats contained in the catalog.

- Another user said that "It would also be interesting to see search interfaces which allow federation across things like Agency data catalogs and so on," as it would allow users to filter and broaden or narrow results with categories. However, a commenter countered, saying, "Federating heterogeneous catalogs is notoriously difficult… Due to this difficulty, it may be better to wait until standard APIs are developed for the Data.gov catalog and then other catalogs could implement that API set to join the federation." As the Data.gov APIs are developed and deployed, users' ability to find and access their desired datasets will improve.

Participants on the IdeaScale site also had some ideas related to improving the accessibility of Data.gov's information to third party search engines (e.g., Google, Bing). While this topic is not directly related to improving the internal search functions of Data.gov, it is relevant to improving the ability of the public to find the data they need. One participant said, "We need a way to better tag and expose data… so that it is effectively crawled by Google and Bing." Similarly, another user noted that Data.gov search will never be better than external search engines, and the focus should instead be on making these external searches work better on Data.gov. The need for a protocol that allows external search engines to intelligently crawl Data.gov is key to helping the public better find data (e.g., OpenSearch, Site Map).

Future releases of the Data.gov website will include a "What's New" section that highlights the latest developments, features, and significant datasets, including a "Dataset of the Day"; a "Data.Gov In the News" section that links to current discussions of Data.gov in the media; and an expanded "Apps Showcase" that contains user-created applications, mash-ups, and widgets that analyze and present data obtained from Data.gov to deliver useful tools to citizens. Eventually, the Apps Showcase will have functionality similar to the Apple App Store[21], with descriptions of the applications, user reviews, and links to the developers' sites to download the applications. Applications will be available for both desktop (Windows, Mac, and Linux) and mobile versions. The current limited version of the App Showcase is shown in Figure 14.

An enhanced search capability in development on Data.gov will include the ability to search by new metadata fields that are not currently available. This may include (not an exhaustive list) how often data is updated or by geographic location, for example.

Another future improvement will be that when users do not find the data they are looking for may submit requests for data, and these requests may be aggregated and shown to the public and Agencies in order to drive the data collection necessary to publish such information.

Since structured data, like datasets and query results, can be related to unstructured documents (like web pages indexed by USAsearch.gov), the Data.gov and USAsearch.gov teams are collaborating on an integrated search API and integrated search box widget that can federate search across both sites and return both structured and unstructured results.

[21] Mention of the site in this document is not an endorsement of the site.

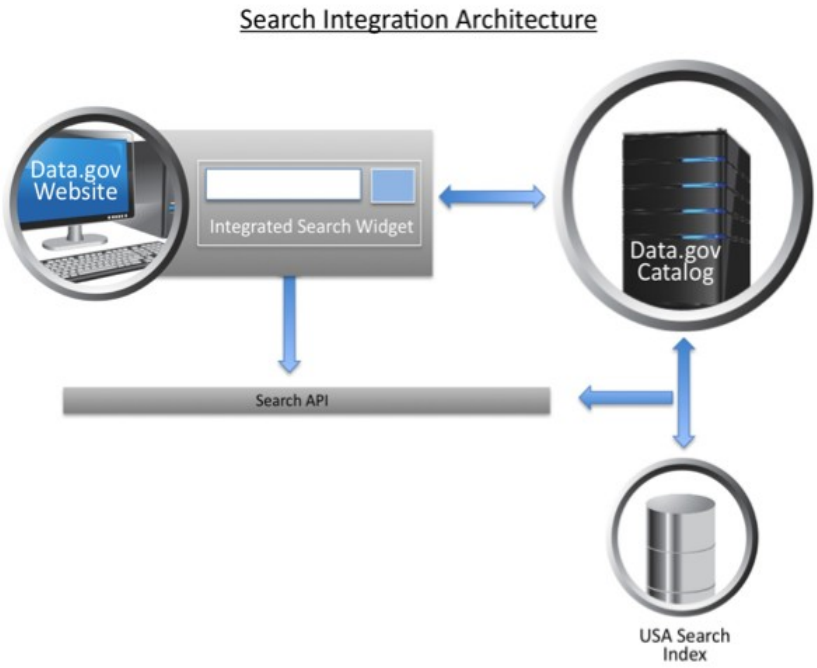

Figure 19: Search Integration Architecture

As depicted in Figure 19, a single integrated search widget can be shared across both the Data.gov and USAsearch.gov websites. This single search widget will use the same integrated search API that will search across both the Data.gov and USAsearch index[22] (it could eventually be federated across other sites).

Other advanced discovery mechanisms including geographic searching are being targeted for future releases. The Data.gov team will work with the FGDC to support their development of this capability for integration into the Data.gov solution. Geospatial search would allow an end-user to draw a bounding box on a map, constrain the results by time or topic areas and then query the Data.gov catalog and visually return the hits that fall into that area. An example of this may be to display icons for any datasets on brownfields in a specific geographic area. In addition to Data.gov implementing this functionality directly, the team is considering expanding the APIs as appropriate to include geospatial search by external websites. Figure 20 is a notional illustration of the concept.

[22] USAsearch uses a commercial search index augmented by an index of editorial and related content.

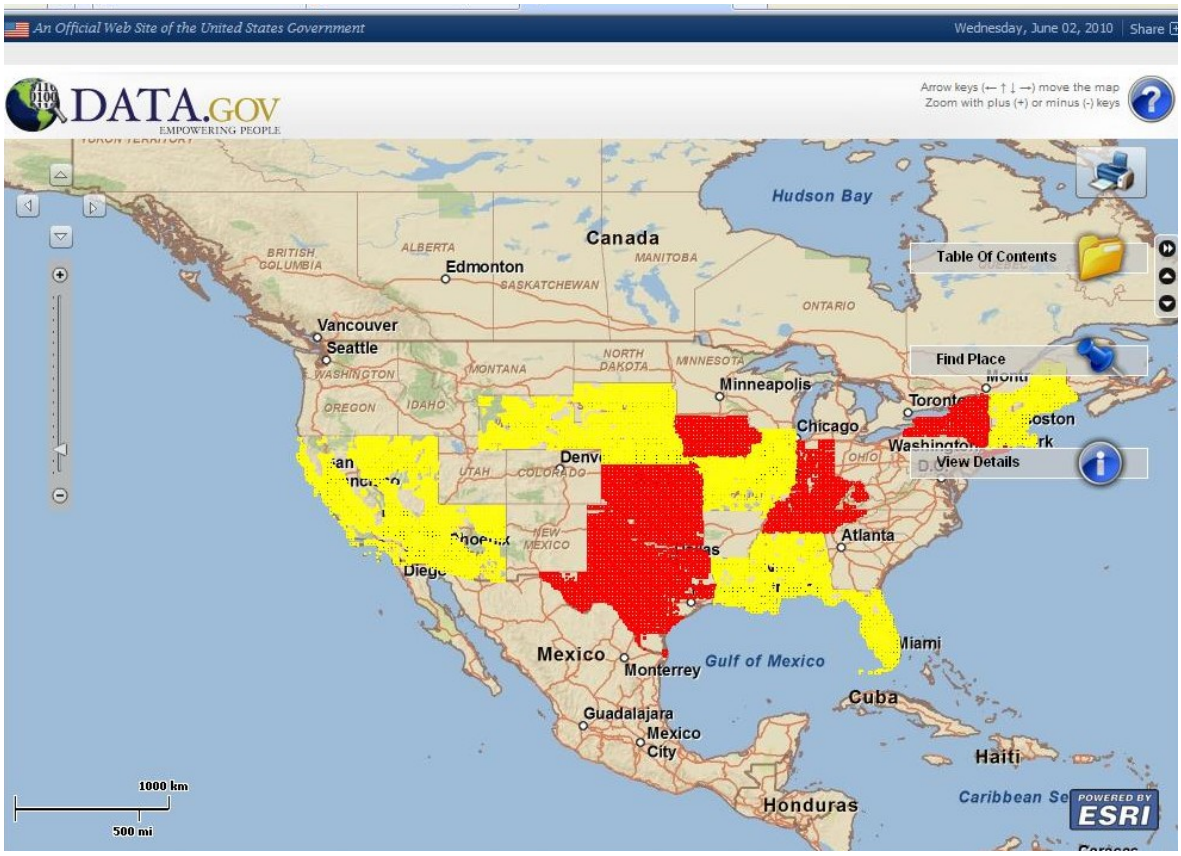

Figure 20: Notional Data.gov Geospatial Search Tool

Another potential tool that would provide substantial value to the public is a geospatially aware search capability that allows users to search for data linked to geographic points of interest. Data.gov has strengthened its partnership with USASearch.gov to develop and deploy this capability. Additionally, the Data.gov PMO team may develop a tool or partner with the relevant organization to tie data and documents that do not contain geographic tags or metadata to specific geographic areas, where they are able to do so.

For example, documents may contain multiple references to specific geographic points of interest such as "Nationals Park". Data.gov is considering tying these to the relevant geographic area, and users performing a geographic search would find such documents and data in their searches. As with the Data-Pedia capability that allows users to define metadata for data, users may also be able to tag data and documents to geographic areas and have other users validate this metadata.

Under this scenario, users may be able to filter geographic searches through multiple criteria, including originating Agency or source, who defined the metadata (user-generated, Agency-generated, automated-generation, or some combination), and other relevant criteria.

3.2.2. How do Agencies Populate Data.gov? Module 2 – The Dataset Management System (DMS)

The Dataset Management System (DMS) was developed to facilitate agencies' efforts to organize and maintain their Data.gov submissions via a web-based user interface. The Data.gov DMS, pictured below in Figure 21, provides agencies a self-service process for publishing datasets into the Data.gov catalog.

The DMS is the approach of choice if an Agency does not have its own metadata repository and does not have the resources to leverage the Data.gov metadata API or harvesting approaches.

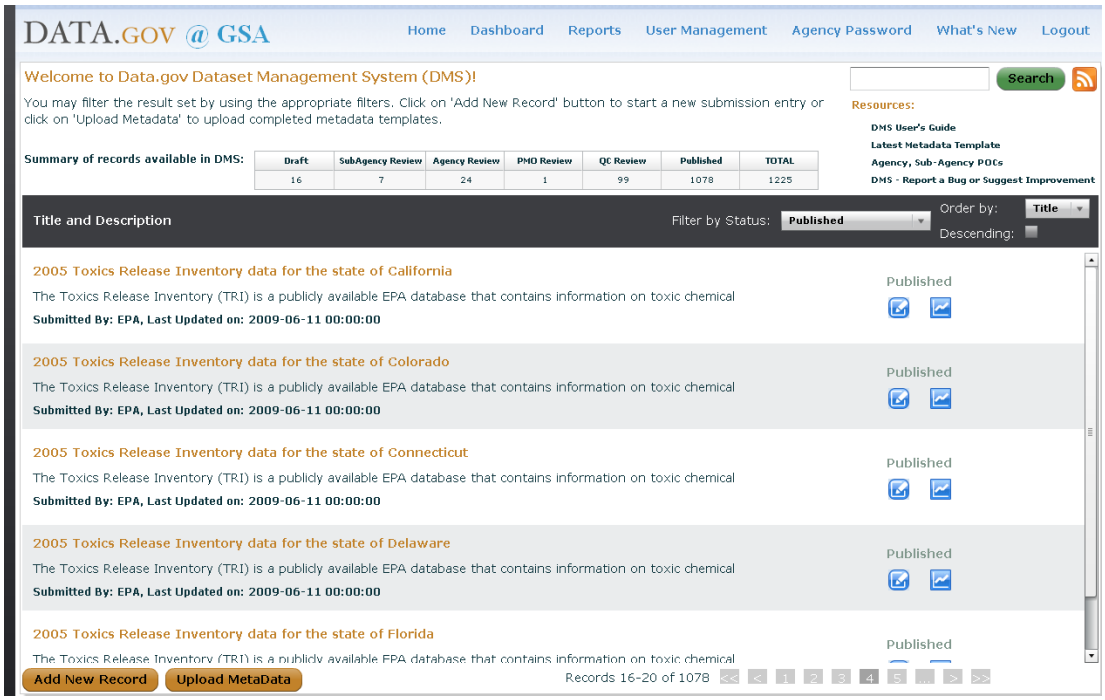

Figure 21: DMS Screenshot

The DMS allows the originators to submit new datasets and review the status of previously submitted datasets. It feeds metadata into the Data.gov catalog on the datasets submitted. New datasets can be submitted either one dataset or multiple datasets at a time. Once a dataset suggestion has been added to the DMS, its status can be tracked through the submission lifecycle, which is shown in Figure 22.

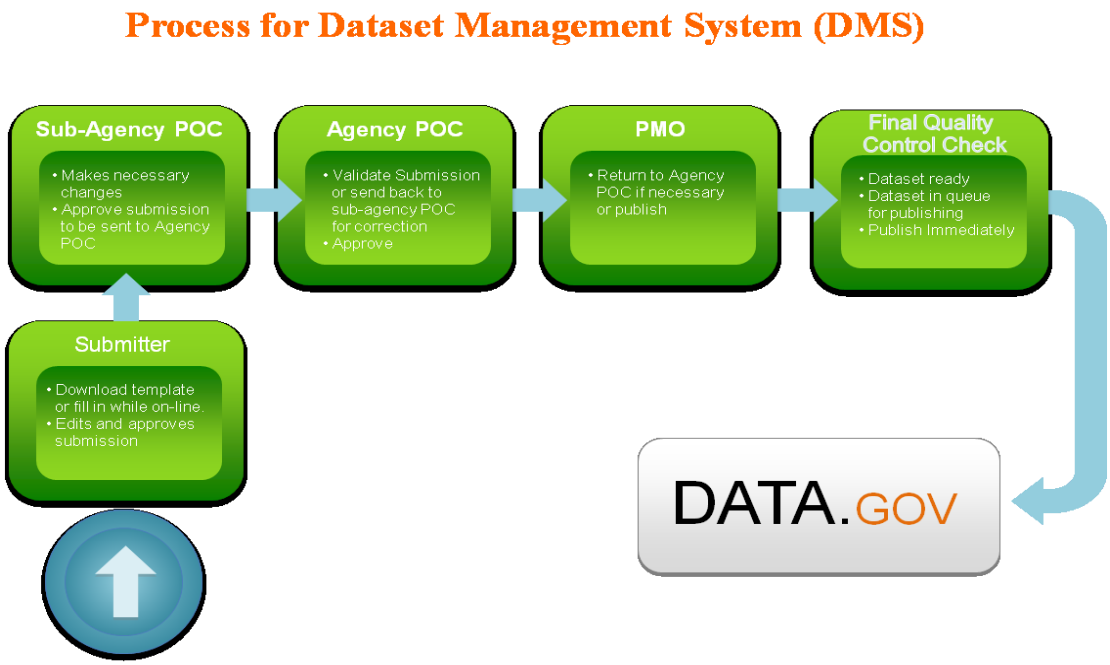

Figure 22: DMS Process Diagram

The DMS enables Data.gov to track the submission and resolution of tickets generated by POCs to provide actionable feedback to agencies. Agency POCs can access the DMS to view the entire published catalog, all published datasets and tools submitted by their Agency, and a dashboard of all pending submissions. The DMS could, in the future, also disclose to the POCs compliance issues that are not being met by the Agency and its data stewards.

A number of the ideas and comments on the IdeaScale site focused are relevant to the DMS, with many participants focused on creating standard definitions and formats and improving metadata quality checks for data available on Data.gov.

- Many participants noted that there was not enough information associated with datasets to be useful or that the metadata was not clear enough to be understood. The addition of standard taxonomies and ontologies discussed in Module 1 would expand the metadata template to make it more clear to users viewing a record or to those searching or browsing for data.
- Users noted problems with the metadata entered into data records. One user suggested that there should be "data quality controls checklist that is completed by the submitter before publishing the data," and another suggested embedding a workflow function into the data management system to track releasability of datasets. A number of quality checks are currently involved in the process of submitting datasets to Data.gov, but specific checklist or checkboxes may be needed at least in the short term to catch inaccuracies as many datasets are added.
- Another idea to improve government data practices was to "use common digital formats for files" so that the public can download and open files easily or so that it will be easier to create mash-ups and integrate data across datasets. Commenters to this idea noted that the Open Government Directive directs agencies to participate in the development and use of voluntary consensus standards and it encourages agencies to consider international data in procurement and regulatory applications. Other IdeaScale participants thought that making data available in as many formats that made sense is the best approach.
- Broadly, participants had lively discussions about the need for creating data standards related to Data.gov or using existing data standards (e.g., Dublin Core, National Information Exchange). One participant said, "Creating standards is hard work and takes a lot of time. There are tremendous benefits to picking up existing global standards rather than creating new ones. Besides, existing standards already have a robust software infrastructure built up around them." Others wondered whether there were specific fields related to open government that should be developed and standardized, or just left voluntary.

3.2.3. Where is the Information About Datasets Stored? Module 3 – The Metadata Catalog and APIs

The metadata catalog is the database that maintains the catalog of datasets and tools listed in Data.gov, along with, as its name implies, the associated metadata. It is the heart of Data.gov, it is how Data.gov knows what datasets are available, what is in them, and where they are hosted. The metadata catalog will evolve into a shared metadata storage service that allows agencies to utilize a metadata repository that is centralized in a Data.gov controlled host, and use it for their own needs. Agencies that do not have metadata repositories of their own will be able to leverage Data.gov's shared metadata repository as a service. So that agencies can leverage the shared metadata repository as an enterprise service, agencies will be able to flag which of their metadata they choose to share with the public via Data.gov

versus those stored in the service but not exposed via Data.gov. Additionally, agencies will be able to designate whether their data contains personally identifiable data and whether the data adheres to information quality requirements.

Figure 23 depicts the key components of a catalog record. It is important to understand that while these various components are drawn in separate boxes, they are actually all part of a single catalog record.

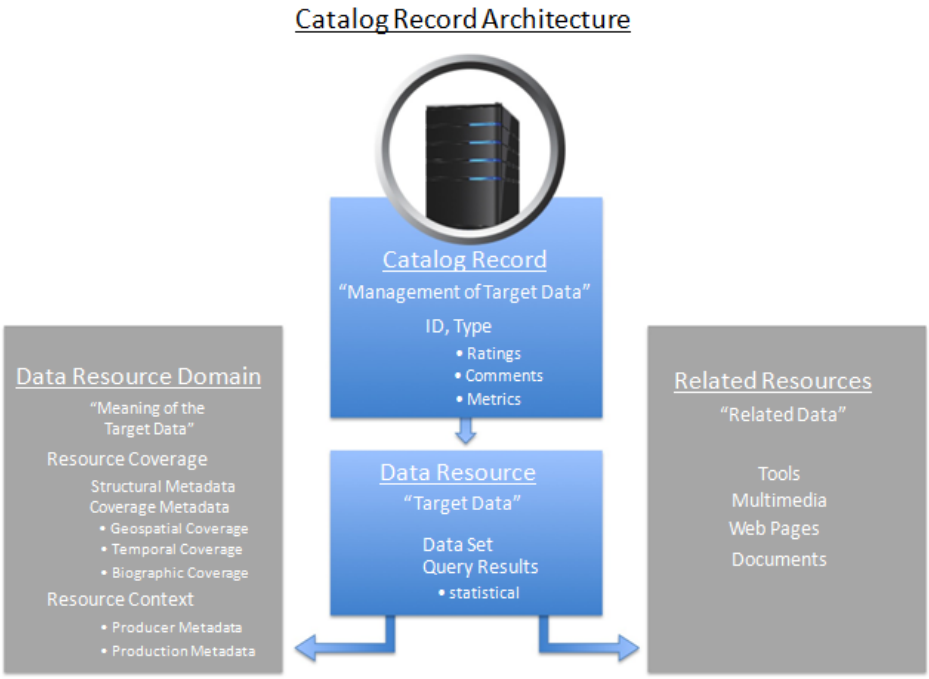

Figure 23: Catalog Record Architecture

The four parts of a robust catalog record are:

- Catalog record header – this part holds both administrative book-keeping parts of the overall record and all data needed to manage the target data resource. To manage a target data resource, this part will keep track of ratings, comments and metrics about the resource.
- Data resource part – a data resource is the target data referred to by the catalog record. A data resource could be a dataset, result set or any new type of structured data pointed to by a catalog record.
- Data resource domain part – a data resource belongs to a domain or area of knowledge. The domain of a data resource has two basic parts: resource coverage and resource context. Resource coverage is a description about what the resource "covers". Resource context is metadata about the environment that produced the data including the production process.
- Related resources part – a structured data resource may have one or more resources related to it. For example, structured data may have images, web pages or other unstructured data (like policy documents) related to it. Additionally, as evidenced on the current site, a dataset may have tools related to it or tools that help visualize or manipulate the data.

As previously noted, developers have been consistently providing feedback that there is an immediate need for Application Programming Interfaces (APIs) associated with

Application developers have consistently asked for APIs.

Data.gov. Based on this public feedback, Data.gov in the process of developing APIs that would provide the ability to interact with Data.gov, as shown below in Figure 24. The APIs would be designed to give programmatic access to the Data.gov catalog entries and the data within the shared data storage service.

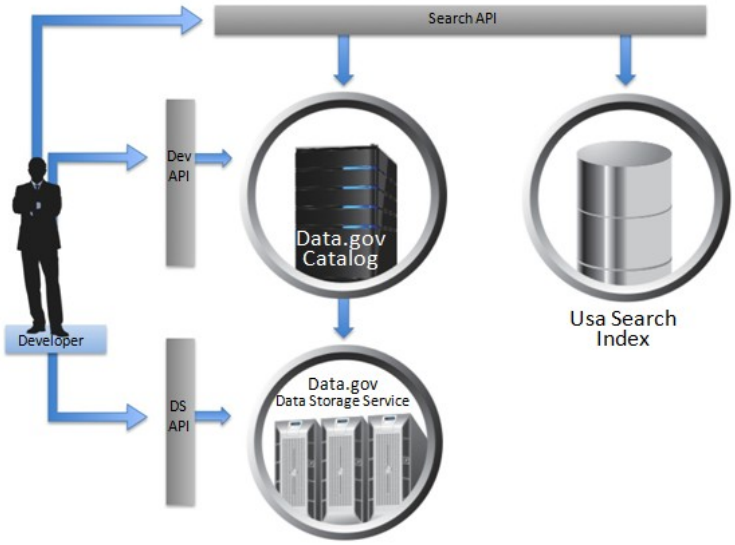

Figure 24: Developer Architecture

Specifically, the APIs under development are designed to be both inbound and outbound. Inbound APIs would allow developers from within the Federal government to submit data or tools to the metadata catalog and submit actual data to the shared data storage service. These inbound APIs would be the most automated way, in the near term, to submit data to Data.gov.

Outbound APIs would allow developers to use the data from the shared data service and the Data.gov metadata catalog to develop their solutions. Developers who build their own websites would be able to leverage the Data.gov metadata catalog or develop their applications using data from the shared data storage services.

As shown in Figure 24, there are three APIs developers use to access specific parts of the Data.gov architecture. The main API, called the DEV or developer API, provides read access to all records in the Data.gov catalog. This includes all components of a catalog record as discussed in the content architecture section below. For searching for datasets and other related data (like unstructured data) the goal is that the developer would eventually be able to use the search API under development to search across both Data.gov and USAsearch.gov. For accessing datasets stored in the Data.gov storage service, the developer would be able to use the data storage (DS) API to retrieve them.

The following are some ideas and comments relevant to developing APIs from the IdeaScale site:

- Participants had some discussion of what kind of web services catalog API would be best for searching and accessing the Data.gov catalog. Users discussed a few approaches, such as Open Publication Distribution System and the Department of Defense Discovery Metadata System (DDMS). One user said, “If Data.gov is reluctant to choose a standard other than CSV, I’d like to see the site consider the creation of a government-wide standard data catalog format by first asking us users to create alternatives for the Data.gov datasets and posting those alternatives through Data.gov for comment…Then, allow for a period of time for comment and even collaboration…That way, Atom or OPDS or something else becomes the core of the ‘universal’

format but 'common' elements can also be defined by two or more government units or verticals. Here Data.gov could provide an incredible service by also pointing to the 'common' extensions. Being optimistic, if Data.gov chooses an approach like OPDS that allows entries to contain sub-catalogs, all of the agencies and even other governments will likely adopt the same approach, especially if the approach is modular, interconnected, and easy to create and extend."

- A user noted that there is a need for common standards *across* government agencies to meet many of Data.gov's goals. The participant suggested that Data.gov work with teams in the major government data centers to develop these standards and offer incentives for their use. Another participant said that having standard APIs across government agencies would help the public integrate data, whether or not they were using Data.gov. The idea of standard APIs across the government is currently part of the CONOPS future vision, but much work is needed to reach that goal.
- One participant suggested that Data.gov support a central metadata system for agencies to use as their own metadata system or to link into to feed the Data.gov metadata system.

3.2.4. How Does Data.gov Keep Track of Everything? Module 4 – Performance Tracking and Analysis

Module 4 performance tracking and analysis engine was an idea generated from public feedback.

As described in Section 3.1.3, Data.gov recently began publishing some of its performance metrics (see Section 1.4) on the website for public viewing. Other metrics are used internally to measure performance and identify potential areas for improvement. Currently, most of the analysis is performed manually. Data.gov is exploring options to include a performance tracking and analysis engine that would be capable of storing information on data dissemination performance. This would provide users with the ability to find, analyze, and report the performance metrics used to track data availability, data usage, and data usability, enhancing transparency and accountability to the citizen. The goal is to combine Data.gov-related measures with Federal-wide data dissemination measures to gain a better understanding of overall Federal data dissemination. Agencies would be the source of the measures to Data.gov and the total set of performance and measurement data would be made available to the public.

A few of the Data.gov IdeaScale ideas and comments saw the value in being able to measure customer satisfaction and use of the site:

- One participant said, "There's no question that open data can provide citizens with more information and transparency about how government is operating and the results it's getting. But that's only half of the equation: We believe that open data produces not just more informed citizens, but also, ultimately, better government. To make that equation come true, though, we have to be vigilant about tracking how people are using data, and incorporating those innovations back into the fabric of government." Similarly, a commenter said that the contributors of the datasets should "receive usage statistics to help demonstrate business needs for the data and for data-sharing tools." Data.gov now provides information on how many times each dataset has been downloaded, and users now have the ability to rate each dataset on overall quality, utility, usefulness, and ease of access. These ratings are available for other users as well as the providing Agency.

- Other participants agreed with the above sentiments, with one saying "Tracking usage is key and any way to capture what the data is supporting can also be tremendously informative." Another, however, pointed out the potential privacy implications of such an approach, asking, "How do you think it's possible to do this without the government being seen as monitoring citizens' use of data? … My sense is that we would have to work out some sort of deliberate way for people to indicate that they voluntarily wanted to participate."
- One participant offered up a framework for tracking performance and meeting the objectives of the OMB Open Government Directive and suggested that it be included in the next version of the CONOPS. The following questions make up this framework:

    1) Are we clear about the performance questions that we want to answer with data to be made available from each of the contributing federal agencies?
    2) Have we identified the availability of the desired data and have we appropriately addressed security and privacy risks or concerns related to making that data available through Data.gov?
    3) Do we understand the burden (level of effort) required to make each of the desired data streams available through Data.gov and is the funding available (either internally or externally) to make the effort a success?
    4) Do we understand how the various data consumer groups will want to see or access the data and does the infrastructure exist to make the data available in the desired format?
    5) Do we have a documented and agreed to strategy that prepares us to digest and respond to public feedback, ideas for innovation, etc., received as a result of making data available through Data.gov?

Performance tracking and analysis will enable Data.gov and citizens alike to see which particular datasets and what kinds of data are providing the most value. By understanding where the value lies, Data.gov can prioritize and focus its efforts on increasing that value.

3.2.5. How Can Data.gov Find More Data to Publish? Module 5 – Data Asset Discovery

Over time, any organization can find that data have been published and exist in the public domain without active management or visibility inside of the organization. The Data.gov team is considering options to assist agencies with identifying previously published data to assist those agencies in their own processes for data management and potential publication to Data.gov. One potential approach, which has recently been successfully piloted, involves deploying a search agent to scan Federal government domains in order to provide data that will assist agencies in evaluating their data management practices and accelerate integration of already public data resources into Data.gov.

An Information Access Tool (IAT) would provide basic capability to identify and characterize data assets that have already been made public on an agency's web site. The goal would be to scan Federal domains and formulate an index of datasets that are already publically available and build reports to deliver to agencies. Associated reporting would serve to provide some basis for the total population of data, provide intelligence to agencies on their potential data assets, and serve to assist the data steward community with an assessment of what is currently exposed to the public.

A recently-completed pilot, conducted in cooperation with the Fish and Wildlife Service (FWS) deployed a commercial IAT to crawl a subset of FWS public website to find datasets that had previously been

published by FWS, but which were not already in the Data.gov catalog. During the pilot, the tool crawled 238 URLs, indexed 863,502 web documents, and found 167 previously unpublished geospatial datasets. These results represent only a fraction of the complete FWS web domain. The pilot established that the search, discovery, and reporting requirements and business logic were sound, realistic, and attainable; found and assessed metadata quality and compliance for geospatial metadata; demonstrated the ability to identify previously unpublished Data.gov geospatial datasets; and executed the first ever successful use of an XML schema to automatically validate geospatial metadata against the Data.gov submission rules and upload the XML file to Data.gov.

The intent of an automated data discovery service is not to automatically populate Data.gov, but rather to assist agencies with their own data inventory, management, and publication processes. The result should be better, more granular agency plans to integrate their already public datasets into Data.gov; more efficient and lower cost data management and dissemination activities through leveraging reported data to jump start and validate data inventories; enhanced ability to develop a proactive understanding of agency compliance with information dissemination and related policy. Most importantly, through continuous measurement the audit tool provides timely and actionable management data to agencies and makes their progress with integration into Data.gov transparent.

3.2.6. How Can Data.gov Make the Data More Accessible and More Useful? Module 6 – Shared Hosting Services

Data.gov is a catalog of datasets available throughout the federal government. The actual data are still hosted by the providing Agencies. In the near future, Data.gov is planning to provide a central data hosting service that combines elements of leading-edge technologies to provide non-proprietary application programming interfaces (APIs) for accessing and browsing the data, social networking to allow discussion about datasets, data virtualization and analytics, onsite mash-ups support, and web publishing. This service would be accessible via APIs and would provide agencies with a cost effective mechanism for storing data that will be made available to the public. The data stored within the service would be made available via feeds and APIs so that the application development community can receive direct enablement from Data.gov. The service's tools enable the application development community and the public to more easily transform the raw data currently available via Data.gov into "finished good" datasets by the public browsing and can be easily consumed by an online community.

The core value proposition to agencies for using the shared dataset hosting service would be to move public download off of their infrastructure, provide better services to data users, and provide data in multiple formats which should reduce total costs and enable more efficient and effective realization of the full Data.gov value proposition.

Providing data conversion into multiple usable formats is as critical as providing the data itself. For instance, the shared hosting service will be used to provide data using query points such as Representational State Transfer (RESTful) web services, web queries, application programming interfaces, or bulk downloads. Data can be made more useful through these services and by extending the metadata template to include data-type-specific or domain-specific elements in addition to the core 'fitness for use' type metadata currently in the Data.gov metadata template. Agency use of query points drives value in some instances. For example, agencies using query points will be able to directly measure "run-time" use of their data as opposed to only recording instances of data downloads. Also, given Agency control over the query point, agencies will be able to better support access to most the current and correct versions of data as well as more clearly understand downstream use and value creation resulting from their data resources.

Data storage and publishing (end user access) would be subject to metering of some sort, to be determined. Given the operational aspect of this module and the need to scale based on volume and end user usage of data, the Data.gov team will look to fully align on the Federal Cloud Computing Initiative and leverage its managed service focus for this module.

Some of the IdeaScale suggestions and comments related to centrally hosting datasets in one place and having tools hosted directly on Data.gov and being able to link datasets easily include:

- One participant said, "Hosting may provide an easier means to agencies for providing public access to data given internal firewalls, infrastructure, and/or budgetary constraints." Related to this idea was a popular idea about reducing duplication amongst federal agencies; the participant said that "Data.gov should become a nationally managed 'access point' that provides a mechanism for all levels of government to participate or integrate with, thus creating a single location for citizens to access government data." This idea is not necessarily about centrally hosting all data, but it could be a step in that direction.
- A few people were also interested in Data.gov serving as the central metadata system host for agencies. As discussed in Module 3, one user said that "Data.gov should support a central metadata system for Agencies to utilize as their own metadata system and that allows Agencies to link their own metadata systems (if they have one) directly to feed the Data.gov metadata systems."
- A number of participants expressed interest in more applications and interactivity that were hosted directly on the Data.gov site, such as exporting tables and graphs, individual pieces of data, or "mashing-up" datasets. While creating this kind of interactivity can be done if datasets are hosted on various websites, it is much more difficult and more likely to run into technical problems that would frustrate users. Users to Data.gov are interested in getting the raw data, but many also want to be able to immediately apply it to other information and look at trends. Data.gov can more easily facilitate these applications if datasets are hosted centrally.

Data.gov recently released an RFQ to award a Blanket Purchase Agreement for the provision of Shared Dataset Hosting. Award of this contract, expected in FY10, will dramatically increase the capabilities available to Data.gov and improve citizens' ability to access, extract, and use the data published by the federal government.

3.3. Looking Forward

The conceptual architecture for Data.gov has evolved in response to the feedback we have received from the public. Like any architecture effort, there are many ways to architect the solution and still satisfy most areas of feedback. The focus in now on delivery of several additional technical components to be based on open standards and aligned with web trends and patterns. Further, the conceptual architecture supports innovation by the Data.gov team and others to increase the number, scope, and operating or business models for data infrastructure tools.

The modular architecture, and the ability to leverage other governmental and other entities with respect to data infrastructure tools, enables the Data.gov team to iteratively build out the solution in line with allocated budgetary resources while still accelerating realization of the end-to-end vision.

3.3.1. Semantic Web

The "semantic web" is a web of data. The original Web mainly concentrated on the interchange of documents. Data, on the other hand, were controlled by applications; each application had its own standards, and they kept data segregated from the Web.

The W3C states that the semantic web is about two things. It is about common formats for integration and combination of data drawn from diverse sources, and it is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing. As the Web of linked documents evolves to include the Web of linked data, Data.gov is working to maximize the potential of Semantic Web technologies to realize the promise of Linked Open Government Data. Data.gov is hosting demonstrations and documents that will help familiarize Data.gov users with this new technology, and that will let citizens and developers work with the government in creating a new generation of "linked data" mash ups. Data.gov now hosts a set of Resource Description Framework (RDF) documents containing “triples” created by converting a number of the Data.gov datasets into this format, making over 6.4 billion triples of open government data available to the community. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a “triple”). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications. An index of all the RDF documents on Data.gov is available at *http://www.data.gov/semantic/data/alpha*. Thanks to our collaboration with the ***Tetherless World Constellation*** at the ***Rensselaer Polytechnic Institute***, Data.gov is now hosting one of the largest open collections of RDF datasets in the world. Other examples of semantic web being used in the government are the U.S. Census Bureau’s DataFerrett (*http://dataferrett.census.gov/*) and the National Library of Medicine’s Unified Medical Language System (UMLS): (*http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html*).

Semantic Web Techniques

The semantic web has a simple value proposition: create a web of data instead of a web of documents. The “web of data” will be designed to be both human and machine readable. The core insight is that data has distinct or overlapping meaning in different contexts. This is a core information technology problem and is manifest in applications such as cross-boundary, cross-domain information sharing, natural language processing, and in enterprise data integration and business intelligence (i.e., mash-ups, dashboards). An example of how this is manifest is the ambiguity highlighted via an example in Wordnet as depicted in Figure 25.

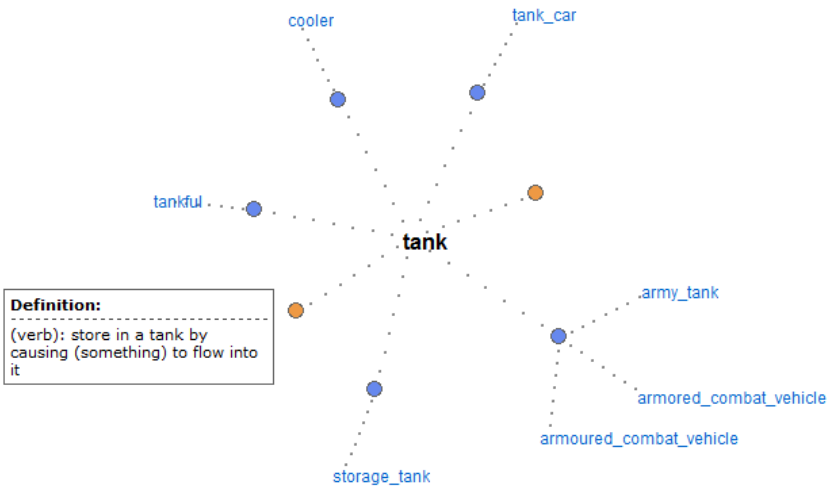

Figure 25: Visualization of Wordnet Synonym Set for "tank"[23]

[23] *http://kylescholz.com/projects/wordnet/* by Kyle Scholz via Creative Commons Attribution 2.5 license.

Figure 25 shows how the word "tank" can have quite a few different meanings as both a verb and a noun. In some applications the context is implicitly understood and this is not an issue. But as soon as two distinct datasets use the same label to have distinct meanings, or the meanings overlap but only partially, or the meanings are the same but that is hidden due to distinct coding or syntactical issues, we introduce ambiguity and most likely defeat the purpose of combining the datasets in the first place.

In order to create this web of data, the W3C and other standards groups have designed specific data modeling techniques to provide such machine readable precision via identification, relationships, advanced modeling and rules. Let's briefly describe each technique and then demonstrate examples of this "curated" data approach. Unique and persistent identification of a unique concept is important to insure unambiguous linking and the accrual of facts on a specific topic. For example, Sir Tim Berners-Lee uses the identifier, *http://www.w3.org/People/Berners-Lee/*, to identify himself and the people he knows using a Resource Description Framework (RDF) formatted data model called FOAF for "Friend of a Friend" as depicted in Figure 26.

Unambiguously identifying all things in a domain is the key first step to enabling machine readable correlation and reasoning about those things. Additionally, by identifying something with a unique Uniform Resource Locator (a URL is a form of URI), one can retrieve a document that provides additional information about the topic and possible equate other things that have been previously identified and are the "same as" this one. Once things are identified, formal relationships between things (and unique identifiers for those relationships) can be asserted. For example, also shown in Figure 26 is the FOAF relationship labeled "knows," which is uniquely identified with the URL: ***http://xmlns.com/foaf/0.1/knows***.

Figure 26: FOAF Visualization[24]

Semantic web modeling expands the traditional modeling techniques of Entity-Relationship Diagrams (ERDs) and Class modeling (as in the Unified Modeling Language or UML) to add powerful logical primitives like relationship characteristics and set theory. Some powerful relationship characteristics are relationships that are "transitive" or "symmetric". A transitive relationship is something like the genealogical relationship "has Ancestor" which is very important in deductive reasoning as is depicted in Figure 27. Additionally, as you can see in the figure, since Matthew "has an ancestor" named Peter and Peter "has an ancestor" named William then it holds that Matthew "has an ancestor" named William.

[24] ttp://foaf-visualizer.org

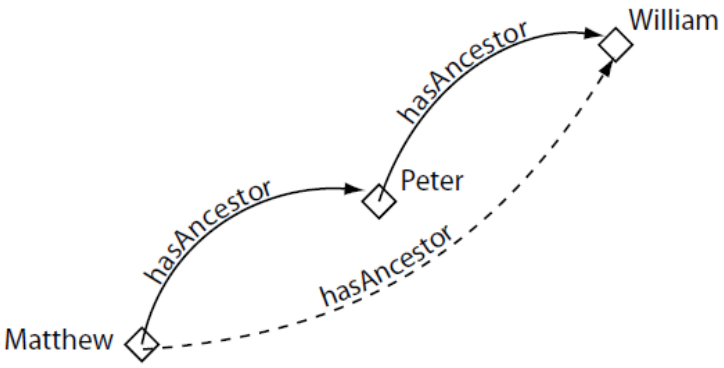

Figure 27: Transitive Genealogical Relationship[25]

A geographic example of a transitive relationship would be "encompasses" as in "Virginia encompasses Prince William County and Prince William County encompasses Manassas". A symmetric relationship is something that holds in both directions. For example, if Mary is "married to" Bill then Bill is "married to" Mary. One final advanced modeling technique is the ability to model types or classes of things using set theory primitives like distinct, intersection and union. This is a very powerful technique for mathematically determining when a logical anomaly has occurred. For example, if a user has an alerting application that is scanning message traffic for the location of a violent criminal on the loose, he/she needs a precise model of a violent criminal as opposed to non-violent criminals (as depicted in Figure 28) and a person cannot be both (or there is an anomaly).

Additionally, to create these advanced domain models there are even free tools, like protégé at *http://protege.stanford.edu*, and many tutorials on the web to educate agencies on these topics.

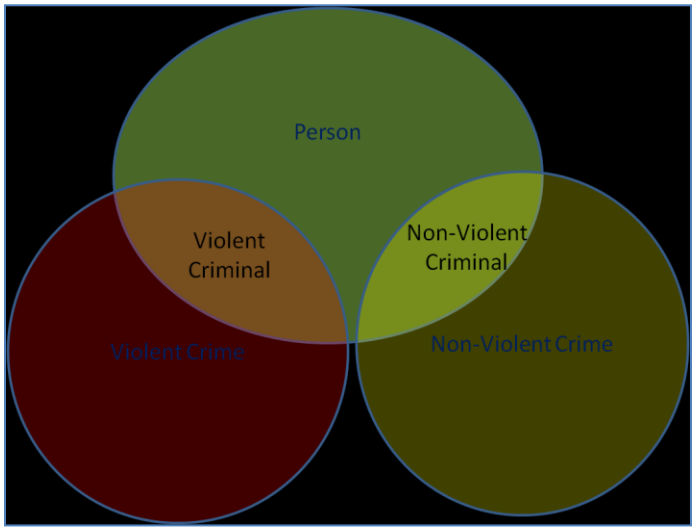

Figure 28: Using Set Theory to Model Violent Criminals

[25]Matthew Horridge, Holger Knublauch, Alan Rector, Robert Stevens and Chris Woe; "A Practical Guide to Building OWL Ontologies Using the Protégé-OWL Plugin and CO-ODE Tools; August 27, 2004; © The University of Manchester; Pg 33.

OMB Memorandum M-06-02 released on December 16, 2005, stated, "when interchanging data among specific identifiable groups or disseminating significant information dissemination products, advance preparation, such as using formal information models, may be necessary to ensure effective interchange or dissemination". OMB Memorandum M-06-02 further noted that "formal information models" would "unambiguously describe information or data for the purpose of enabling precise exchange between systems".

The government has also produced several cross-domain data models that can be leveraged to improve both semantic understanding and discoverability of government datasets. The *National Information Exchange Model (NIEM)* and the *Universal Core (UCore)* are two robust data models that are gaining traction, incorporating new domains and increasing information sharing across federal agencies, the Department of Defense and the Intelligence Community. The NIEM data model is designed in accordance with Resource Description Framework (RDF) principles and can generate an OWL representation. NIEM has extensive use across levels and domains of government. In particular, it has been endorsed by the *National Association of State Chief Information Officers*. The US Army has created the UCore-Semantic Layer (SL) that is an OWL representation of the basic interrogative concepts (who, what, when, and where). These efforts are prime examples of the government's ability and commitment to providing robust tagging and modeling mechanisms to improve discovery of, sharing of and eventually reasoning about federal data.

Today's "industry best practices" are more frequently grounded in semantic techniques that enable the semantic web and query points that the public can directly access (like Amazon Web Services[26]). Under this model, it is the (formally coded) data concepts themselves that are cross-linked, as opposed to just cross linked web pages. There is a push among some search engine companies to create standards for indicating certain kinds of metadata directly within web pages. Rich Snippets from Google and Search Monkey from Yahoo[27] are competing attempts (but with similar goals) to allow content developers to associate structured data with information shown on their websites. They currently support a variety of formats, including micro formats and Resource Description Framework (RDF).

In accordance with the philosophy of OMB Memorandum M-06-02, and leveraging today's mainstream "formal information model" capabilities, the evolution of Data.gov will include the incorporation of semantically enabled techniques within the sites and within the datasets themselves. The thought work for this progression has already begun and has been aided by feedback from the public.

Curation is the process of selecting, organizing and presenting the right items in a collection that best deliver a desired outcome. Curation of data is preparing data so that it is more usable and more exploitable by more applications. In that light, the semantic web techniques previously discussed are the next logical step in the widespread curation of data. In particular, it is a leading edge, potential best practice in Federal data management.

[26] This reference is an example and not an endorsement.

[27] These references are examples and not an endorsement.

Figure 29: Computing Results from Curated Data[28]

A good example of the benefits of such curation is the Wolfram Alpha website (*http://www.wolframalpha.com*). Wolfram Alpha exclusively uses curated data in order to calculate meaningful results to queries. For example, returning to our crime scenario, a user could input to Wolfram Alpha, "violent crime in Virginia/violent crime in the US" and it computes the information in Figure 29.

Other benefits of using semantic web techniques include cross-domain correlation, rule-based alerting and robust anomaly detection. While out of scope for this document, it should be obvious that increasing the fidelity of data increases its applicability to solving problems and increases its value to the Data.gov developer and end-user.

The Semantic Web Roadmap

Semantic web techniques are not yet widespread in the Federal government. Given our principle of program control, Data.gov takes an evolutionary approach to implementing these techniques. Such an evolution involves pilots, a piece-meal transition and a lot of education. The result will be to demonstrate the value proposition, establish end user demand, and empower data stewards to adopt semantic web techniques. In order to accelerate evolution, an experimental semantic-web-driven site will be established as depicted in Figure 30.

[28] The inclusion of this screenshot is not an endorsement of the site.

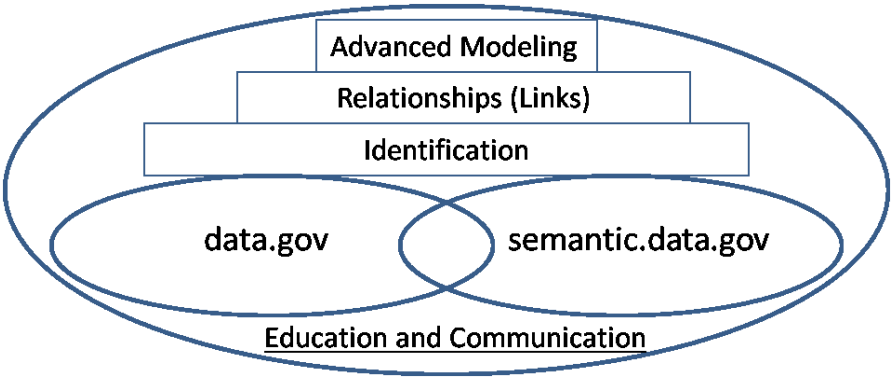

Figure 30: Semantic Evolution of Data.gov

In addition to Agency pilots, the semantic.Data.gov site will leverage lessons learned from the United Kingdom's version of Data.gov (soon to be released) which will be built entirely on semantic web technologies. An ancillary benefit of piloting techniques like unique identification and explicit relationships is that the lessons learned will assist the more traditional implementations of these techniques on Data.gov. It is envisioned that as the benefits and applications based on semantic Data.gov datasets increase, a migration and transition plan will be developed to merge the efforts.

The evolution of Data.gov will include a progression towards the semantic web, a fast moving space that will fundamentally transform the web. It is expected that the UK version of Data.gov will be using a semantic web approach. The U.S. Library of Congress is a best practice example of a Federal organization that is already moving towards the semantic web with its *"Authorities and Vocabularies" service*. Agencies can study approaches like that used by the Library of Congress in anticipation of semantic.data.gov. An Agency that owns/defines authoritative domain data will eventually be asked to put the domain specifications (metadata) and the corresponding instance data on the web using semantic techniques. Working groups under the Senior Advisory Council will focus energies on establishing the relationships (the links) between these authoritative datasets. In some instances, relationships may already exist and simply need to be adopted by the data stewards.

3.3.2. Geodata Integration

Geodata represent the vast majority of datasets available through Data.gov. Many of the geo datasets originated at Geodata.gov, a geographic information system (GIS) portal (also known as the Geospatial One-Stop (GOS)), that serves as a public gateway for improving access to geospatial information and data under the Geospatial One-Stop E-Government initiative. Geodata.gov was launched in 2004 to facilitate communication and sharing of geographic data and resources to enhance government efficiency and improve citizen services by making it easier, faster and less expensive for all levels of government and the public to access geospatial information.

The portal is a catalog of geospatial information containing thousands of metadata records (information about the data) and links to live maps, features, and catalog services, downloadable data sets, images, clearinghouses, map files, and more. The metadata records were submitted to the portal by government agencies, individuals, and companies, or by harvesting the data from geospatial clearinghouses.

One user commented to IdeaScale: "Geographic referencing adds critical context to data. It helps users quickly and easily determine whether a dataset pertains to their specific area of interest, and in the event that it does, empowers users by immediately allowing them to visualize that data, perhaps coupled with additional datasets for informing context. Both Geospatial One Stop and Data.gov are

citizen centric initiatives. Migrating and consolidating the two programs would both energize and maximize any place based analytical capability the nation could leverage in the future."

Plans are underway to merge geodata.gov into Data.gov to eliminate the duplication of their functions and solidify Data.gov's status as the single, central source for finding and obtaining government data, including geospatial data.

In addition to the upcoming merger of Geodata.gov into Data.gov, a new feature that enables users to preview and visualize geospatial information, and create mash-ups with other data from Data.gov, basically creating an annotated map in real time. Examples of visualizations are shown in Figure 31.

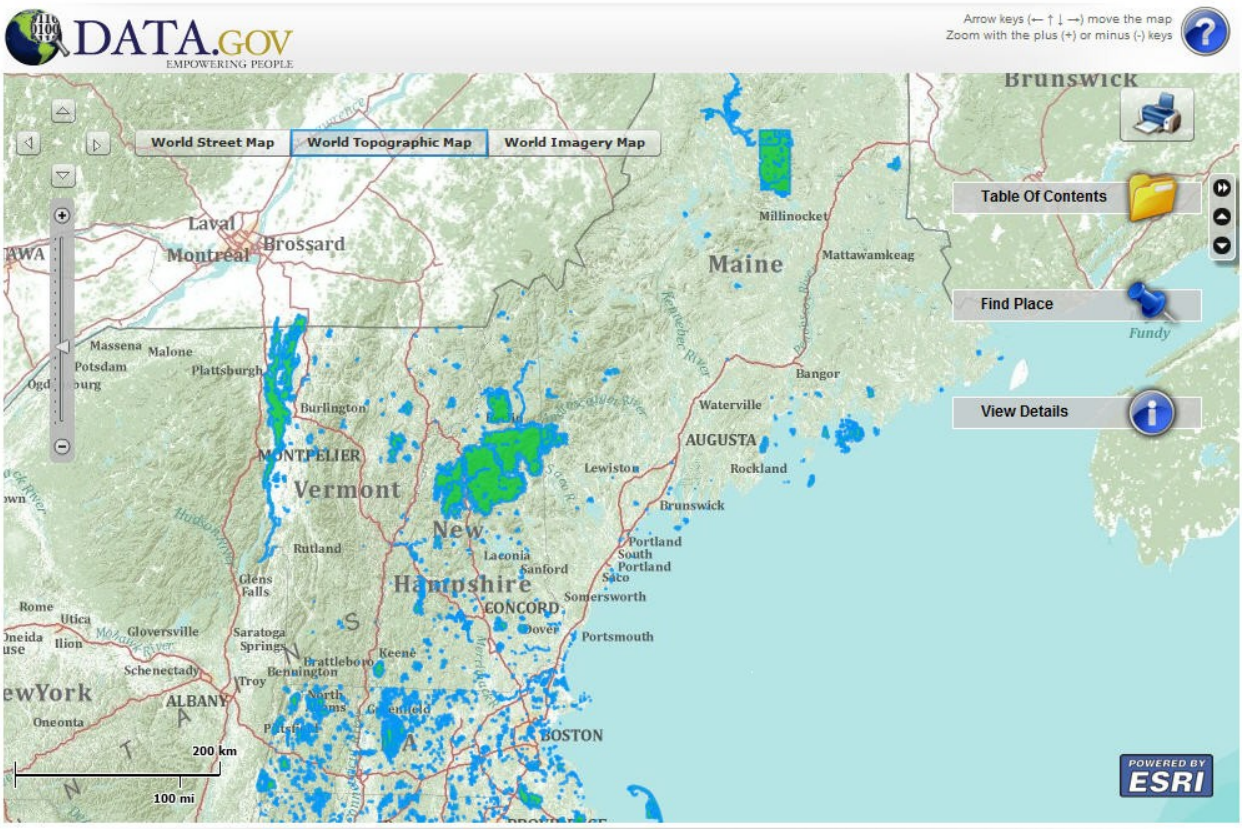

Figure 31: Geodata Visualization Examples

Geo-enabling Data.Gov involves integration of Geodata.gov into Data.Gov and the incorporation of next generation data Geo-visualization and mash-up capabilities. The Geospatial search module of Data.Gov would allow an end-user to draw a bounding box on a map, constrain the results by time or topic areas and then query the Data.Gov catalog and visually return the hits that fall into that area.

As requested by the public, Data.gov visualization services could be delivered through the site and could include analytics, graphics, charting, and other ways of using the data. In many cases enhanced visualizations will be delivered by the Data.gov team or others as data infrastructure tools, built on top of published APIs. These enhanced visualizations or other uses will in some cases be accessed via the Data.gov site and in others via external web sites.

3.3.3. Communities

One IdeaScale commenter made the suggestion "One way to envision the contributions of Data.gov is to build a series of scenarios, maybe one for each major issue that if left unresolved could lead to a major disruption." A Community is similar to a category, or a scenario, in that it is a theme that runs through a large set of datasets, crossing organizational and functional boundaries.

The first Data.gov community to deploy was Data.gov/RestoreTheGulf, a page dedicated to providing data related to the response to the Deepwater Horizon oil spill (Figure 32). Data include oil and gas flow and recovery measurements, air and water sample data, oil spill-related exposure information, and other data of interest to scientists, recovery workers, and citizens, as well as links to state and local sites where additional oil-spill related data can be found.

Figure 32: Data.gov/RestoreTheGulf

Data Communities may be permanent, such as Health and Education, or they may be short term, like RestoreTheGulf. What all communities have in common is that they provide citizens with easy access to data that are relevant to their particular areas of interest.

3.3.4. Data-Pedia

Data.gov's goal is to be the central source for government information. Its ability to do so, especially in regard to collaboration with state, local, and tribal governments, rests upon the perceived value of the tools it provides to the general public and other groups. One tool being considered to enable this goal would be a service that allows users to integrate locally gathered and stored data with other similar data to create comprehensive, national level datasets. For example, municipalities, local governments, property developers, and other real-estate focused groups collect and store property map information for vast areas of land throughout the US. These maps, however, are not centralized in one location, and users often have to navigate to each county's or city's website to find property map information. By enabling users to upload such information, Data.gov could act as a central repository that investors, the general public, and others may use as they see fit. As this tool might grow and evolve, it may be possible to show changes in property lines over time as areas are developed and urbanized.

Data-Pedia may also serve as a tool that allows the public to add a tag to a dataset's metadata. Similar to the way Google and Flickr utilize the public to tag photographs, Data.gov could empower users to tag metadata and allow citizens to upvote the most relevant metadata submissions for data. This would both improve the quality of the metadata and collaboratively engage the public to extend Data.gov's resource base by crowdsourcing certain functions.

Together, these two capabilities would bring a useful tool with clear value to the public, the Federal government, and others. Moreover, this would enable stakeholders from outside the Federal government to lead initiatives and utilize Data.gov as a central collaboration and coordination tool, thus leaving the strategic direction of any initiatives within Data-Pedia to their creators and collaborators.

3.3.5. Mobile Applications

The public continues to expand the ways in which it accesses information, including via mobile devices. The success of the iPhone and Android application markets indicates Data.gov's need to provide mobile services and infrastructure that will enable users to access information via smart phones. The website showcases a variety of mobile apps that have been developed by citizens. In addition, Data.gov may create mobile applications and provide the necessary APIs, as described in Section 3.2.3, to allow independent developers to create third-party applications. This would increase Data.gov's mobile footprint and reach by providing the necessary infrastructure for independent developers to disseminate information aggregated by Data.gov.

3.3.6. Agency and Site Performance Dashboards

The Agency and site performance dashboards display the relevant metrics that are collected by the performance and analysis engine. As previously discussed, each Agency may collect and share performance metric information with Data.gov through an automated process. This process would standardize the incoming performance data, and then load the data into a viewable dashboard environment that could be displayed to the public, Data.gov personnel, and Agency personnel. The public's performance dashboard would have limited access to the performance metrics. The performance data would be re-usable across Federal websites as well as by the public.

3.4. Working with Other Government Websites

Data.gov also can be seen as the source location to access structured data behind some of the government's most significant websites. Existing and newer websites such as *USA.gov*, the *Federal IT Dashboard*, *USASearch.gov*, *FBO.gov*, *USAspending.gov*, *Geospatial One Stop*, *FedStats*, and *Grants.gov* all have major presentations of data using search and presentation technologies. The structured data

behind these websites will be part of the inventorying and metadata harvesting process. These other initiatives are expected, like the agencies, to register their data and tools with Data.gov so that Data.gov includes the most appropriate inventory of data and tools available to the public. Data stewards who previously published to these sites may continue to do so as these sites, once they register their data and tools with Data.gov, will be integrated with the Data.gov solution. Additionally, any of these sites that require reports from agencies should also move to require reports in machine-readable formats.

Agencies that have geospatial data are in many cases publishing that data to Geospatial One Stop (GOS) today. The harvesting process used by GOS is mirrored in the conceptual solution architecture described above and points to a roadmap for further integration. The Data.gov team will work with the GOS team to pursue further integration of GOS into Data.gov.

In addition to working with other Federal agencies and initiatives, the Data.gov team is working with the National Association of State CIOs (NASCIO) to share standards and arrive at compatible concepts of operation. The Data.gov PMO will look to expand similar relationships in the US and internationally. These relationships may be modeled on the formal structure that OMB and the Data.gov team are using to engage and establish a long-term collaborative relationship with other federal entities.

Data.gov Concept of Operations