

“Transparency, Accountability, and Oversight”

Data.gov Concept of Operations



Office of E-Government and IT
Office of Management and Budget



*Powered by the Federal Chief Information Officers Council
and the Federal Enterprise Architecture*

Version 0.93

September 3, 2009

Contents

1.	Data.gov Strategic Intent.....	1
1.1.	Data.gov Vision and the Open Government Agenda	3
1.2.	An Illustration of How Data.gov Provides Value to the Public	5
1.3.	Value Proposition to Agencies.....	6
1.4.	Measuring Success	9
2.	Data.gov Operational Overview	10
2.1.	Functionality of Data.gov	10
2.2.	How Datasets and Tools Get to Data.gov.....	12
2.3.	Hosting Data.gov Using Cloud Services	16
2.4.	Who Makes Data.gov Work: Key Partnerships.....	17
2.5.	What Enables Data.gov: The Dataset Management System	21
3.	The Path Forward	22
3.1.	Enhanced Requirements	22
3.2.	Modernizing Agency Data and Metadata Publishing	23
3.3.	Assisting Agencies with Discovering Previously Published Data	27
3.4.	Working with Other Federal Websites	27
3.5.	Incorporating the Semantic Web	27
4.	Transitioning to the Future.....	29
4.1.	Step 1: Establish Enhanced Governance	29
4.2.	Step 2: Issue New Standard Guidance and Reinforce Existing Requirements to Agencies.....	29
4.3.	Step 3: Sequence the Implementation of the Enhancements	31
5.	Reference.....	32

1. Data.gov Strategic Intent

Data.gov is an Administration initiative intended to allow the public to easily find, download, and use datasets that are generated and held by the Federal Government. Data.gov was launched on May 21, 2009, emphasizing the President's commitment to open and transparent government. Data.gov is intended to be the Federal Government's marquee brand for reliable, authoritative data that adhere to information quality guidelines and can be leveraged by the public as the Federal Government's preeminent collection of data. Establishing Data.gov as a marquee brand requires that agencies provide their most relevant and informative datasets and tools for presentation through Data.gov. Establishing and maintaining the marquee brand require that the scope of Data.gov be kept to Federal datasets that Federal data stewards can validate as authoritative and compliant with information quality policies.

The Department of the Interior's and the Environmental Protection Agency's Chief Information Officers (CIOs) serve as the co-leads for Data.gov, with support provided by the General Services Administration and the Office of Management and Budget. This team worked with many agency points of contact to launch Data.gov on May 21, 2009. The Data.gov launch signaled the President's commitment to open and transparent government. The launch of Data.gov also was a continued implementation of the [M-06-02 memorandum titled "Improving Public Access to and Dissemination of Government Information and Using the Federal Enterprise Architecture Data Reference Model"](#)¹. Specifically [OMB Memorandum M-06-02](#) requires agencies to "organize and categorize [agency] information intended for public access" and "make [data] searchable across agencies". Data.gov is a major enabler for agencies to fulfill requirements in [OMB Memorandum M-06-02](#) in a more consistent and citizen friendly way.

Improving the accessibility of the massive data stores of Federal Government data can help foster innovation, new discoveries, and fuel the knowledge economy of the United States and the world. Because Data.gov is a focused effort on "open data and tools", it sets a new direction for the transformation of government data through the next generation of information interoperability and the exchange of ideas and innovation by an empowered public, industry, academia, and all levels of government. This bold vision in this information age will spur advancements in the Nation's information agility thereby helping us solve problems in a more collaborative and effective manner. Relevant, accurate, and timely data sources will also help us make more informed decisions and collaborate with common frames of reference, which will engender increased trust between the government and its stakeholders.

To achieve the intent of making Executive Branch data more easily discoverable and useable, implementation of Data.gov will include an emphasis on the development of internal guidance and processes that facilitate data sharing to the extent permitted by law.

¹ [OMB Memorandum M-06-02](#) 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."

Implementation goals include:

- Enabling transparency at the authoritative source of record and enabling the agencies to manage the metadata and publication frequency of their datasets.
 - Improving the way that Federal datasets are described so that the datasets are discoverable. This includes facilitating uniformity in the provision of metadata (contextual information about the data).
 - Promoting authoritative data sources within the agencies. This includes leveraging authoritative data sources to build the Data.gov data catalogs, and ensuring that authoritative data sources are published in appropriate machine readable formats.
- Increasing the consistency of information dissemination efforts across the Federal Government, as well as improving compliance with existing statutory responsibilities and Executive Office guidance regarding information quality, accessibility, privacy, confidentiality, security, transparency, and related concerns.
 - Encourage transparent measurement and reporting by leveraging the Federal Enterprise Architecture Assessment Framework and the e-Gov related dashboards. Agencies will be measured on their overall compliance with existing information requirements.

The key implementation strategies for Data.gov across the Executive Branch include:

- Ensuring that the structure of Data.gov enables the public to have a one stop point of entry to find data and tools by providing standardized, harvestable metadata and effective search and discovery mechanisms.
- Establishing an engaged Executive Branch community in which Chief Information Officers designate “Points of Contact” (POCs) to coordinate with the data stewards inside their agencies to identify datasets and tools for inclusion on Data.gov.
- Enabling each agency to develop the internal process that works best for identifying submissions to Data.gov, including managing relationships with the programs that own or house the agency’s data assets (managing relationships with data stewards and obtaining program-level sign off of dataset submission) and ensuring conformance with existing statutory responsibilities.
- Creating a learning environment among agency Data.gov Points of Contact so that the latest advances in best data practices including search and discovery are shared among agencies.
- Streamlining the management and operations of Data.gov by allowing agencies to self manage providing content for which they are responsible.
- Streamlining Data.gov feedback mechanisms between the public and sponsoring agencies with the intent to limit management layers so that feedback is delivered in a timely manner and in an actionable format.
- Establishing a performance management framework for information dissemination with agencies to allow for consistent and cross government measures of value creation.

- Leveraging the Federal Enterprise Architecture (FEA) and related frameworks, methodologies, standards, and assessment frameworks to further align, guide, and coordinate agency activity.

On a broader level, implementation of the Data.gov Project Management Office will include:

- Coordinating metadata and collaborating actively with States and local governments launching their own Data.gov sites with the strategy of making access to data among levels of government increasingly seamless for the public.

1.1. Data.gov Vision and the Open Government Agenda

The vision is that Data.gov provides improved access to high quality government data and tools to the public. [OMB Circular A-130](#) states that “the free flow of information between the government and the public is essential to a democratic society”. Datasets delivered through Data.gov are intended to be authoritative in nature, actively managed by dataset stewards, presented with the appropriate metadata to be easily discoverable. Datasets delivered through Data.gov are also intended to be published in format(s) that are easily useable including multiple file types and query points² such as application programming interfaces, web services, and web query methods of data access.

Data.gov is a catalyst and focal point. The Federal Government faces the challenge of how to be open and transparent in a world of emerging web capabilities. The Nation’s public sector ability to generate information has historically outpaced the ability to easily find and reuse information. Data.gov is working the “find and reuse problem” within the OMB information dissemination policy framework.

The vision of more broad and accessible data dissemination is grounded in the [Paperwork Reduction Act of 1995](#)³ which lays out statutory requirements for agencies with respect to information dissemination and, among other things, requires to the extent feasible, that each agency should, when disclosing information in electronic format, provide timely and equitable access to the underlying data. The vision for broad dissemination of data is further grounded in the [E-Government Act of 2002](#)⁴ which requires

² The [Federal Enterprise Architecture Data Reference Model](#) defines a query point as “An endpoint that provides an interface for accessing and querying a Data Asset. A concrete representation of a Query Point may be a specific URL at which a query Web Service may be invoked.”

Query points are agency hosted web services

³ Section 3506 of the Paperwork Reduction Act states “With respect to information dissemination, each agency shall ensure that the public has timely and equitable access to the agency’s public information, including ensuring such access through (A) encouraging a diversity of public and private sources for information based on government public information; (B) in cases in which the agency provides public information maintained in electronic format, providing timely and equitable access to the underlying data (in whole or in part); and (C) agency dissemination of public information in an efficient, effective, and economical manner”

⁴ Section 207(d) of the E-Government Act 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.”

OMB to issue policies requiring “the adoption of standards, which are open to the maximum extent feasible, to enable the organization and categorization of Government information ... in a way that is searchable electronically, including by searchable identifiers”.

The vision for Data.gov is further reinforced in the President’s January 21, 2009 [*Transparency and Open Government Memorandum*](#)⁵ where he states that information should be disclosed “rapidly in forms that the public can readily find and use.” The President continues by stating that government should be “transparent”, “participatory”, and “collaborative”.

Data.gov will help implement the President’s open government agenda. Transparency, participation, and collaboration are the three pillars of the open government agenda and each is facilitated by a successful Data.gov.

Transparency

At the core of Data.gov is the intent to make Federal sector datasets more transparent. By increasing the ability of the public to discover the vast stores of government data, we increase transparency. Dissemination of public domain data has always been an integral mission activity. Data.gov takes this traditional activity to the next step by providing coordinated cross-agency access to datasets via a non-agency specific delivery channel.

Data.gov enhances the ability to find information by offering a catalog of datasets across agencies. The objective is that the public should not have to have knowledge in advance of how Federal agencies and data programs are organized in order to find datasets about a given topic or determine the characteristics of a known dataset. Furthermore, Data.gov is developing search capabilities that will enhance the public’s ability to find very specific pieces of information within those datasets. Data mining and extraction tools offer the public another approach to find specific statistics, especially when they do not have the skills to analyze the underlying data or the underlying data cannot be made available to the public due to security or confidentiality concerns. Such accessibility also enhances the ability of Federal agencies to more easily discover information available from other agencies for common mission purposes.

Participation

Public participation is a key pillar of the open government agenda and is critical to the success of Data.gov. Data.gov is both an enabler of public participation, as well as a recipient of the benefits of public participation. Data.gov enables the public to participate in research and knowledge discovery by providing Federal datasets from which they can build applications, conduct analyses, and perform research. The value proposition of Data.gov is that Federal datasets can be used by the end user community to better advance discovery and knowledge creation.

This Administration is particularly interested in developing ways for the public to become more engaged in the governance of our Nation. One approach is to encourage the public to analyze issues from their own perspective with the data they have paid for with their tax dollars. Data.gov seeks to engage the

⁵ The President’s Open Government and Transparency memo states “Government should be transparent. Transparency promotes accountability and provides information for citizens about what their Government is doing. Information maintained by the Federal Government is a national asset.”

public in expanding the creative use of Federal data beyond the walls of government by encouraging the development of innovative ideas (e.g., web applications); combining Federal and other datasets to gain new insight; pursuing new economic and socially-based ventures; and thereby enriching the lives of citizens.

To enhance the ability of programmers and developers to create new capabilities for the public with minimal investment by the government, Data.gov needs to foster the availability of data in a form compatible with the next wave of the Internet -- the semantic web or data web. Protocols are being developed which encode meaning into datasets in such a way that they are directly interpretable by computers, so that instead of having “data on the web”, there will be a “web of interoperable data”. Information has become mobile; no longer limited to those who know how to access a particular system. Through the data web, data aggregation and analysis can be done instantaneously, and new applications and services can be created. The Federal Government has a wealth of information; if made more available in accessible formats, an explosion of innovation can take place. Data.gov is designed to foster this paradigm shift.

As such, Data.gov places a strong emphasis on the dissemination of public information generated by the Federal Government in platform-independent standardized formats that promote creative analysis – via data that is authoritative and granular.

Collaboration

A key mechanism to improve and evolve Data.gov is direct feedback, comments, and recommendations from the public. For example, individuals are encouraged to suggest datasets to add to Data.gov, rate and comment on the value and quality of current datasets, and suggest ways to improve the site overall. The agency Chief Information Officers’ points of contacts (POCs) receive the applicable public suggestions so that agencies can respond accordingly.

1.2. *An Illustration of How Data.gov Provides Value to the Public*

An important value proposition of Data.gov is that it allows members of the public who have the ability to leverage Federal datasets for a robust discovery of information, knowledge, and advancement. We can maximize the return on investments in generating Federal datasets by breaking down agency stovepipes, providing data in reusable formats, and facilitating a standardized approach to metadata elements that enable increased search abilities. As was the case for the Human Genome project, releasing datasets beyond the walls of government allowed for public access, facilitating creativity and ingenuity. Understanding the potential value rests with considering the nature and quantity of the Federal datasets themselves.

The amount of Federal data that is generated and managed across the government and made available to the public is substantial. The total number of datasets is unknown and depending on the definition used is likely to be in the millions. The wide array of datasets is significant but the quantity of access points for the datasets is startling. Figure 1 illustrates that there are thousands of Federal Government websites that exist within roughly 2,500 registered Federal domains. The public’s search for authoritative, high quality data is frustrating as the data are fragmented across the organization-specific websites of the entire Federal Government.

End Users of Data.gov

The user community for Data.gov is not a single person group, but rather a collection of groups including data enthusiasts, technical developers, visualization experts, oversight organizations, researchers, academics, media, mission advocates, and individuals interested in knowledge discovery. The user communities are all encouraged to continue providing direct and indirect feedback to Data.gov. Direct feedback takes the form of email and comment and rating postings through the website. Indirect feedback includes blog postings, tweets, magazine articles, conference panel discussions, and both traditional and new approaches to engaging in the Data.gov “conversation”.

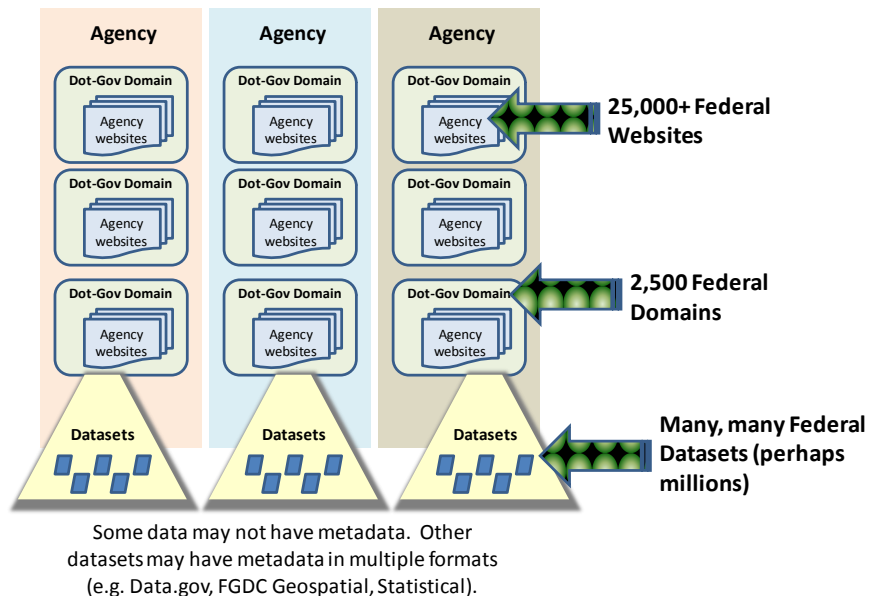


Figure 1: The Scope of Federal Websites

Data.gov exposes Federal Government data at the source by providing a searchable catalog of metadata that describe the underlying data. In order to assist in cataloging different metadata standards across multiple agencies, Data.gov established a standard metadata template. This template was developed by extending the existing Dublin Core metadata standard. The public receives the benefit of searching Federal datasets from one location, and searching across metadata that are standardized across the Federal Government. The public also receives the benefit of linking directly from the standard metadata to the actual downloadable datasets, the top-level agency domains, and the agency program websites that serve as the actual source for the data. In short, Data.gov provides the public with a single access point for Federal datasets, the ability to search standard metadata across those datasets, and the ability to provide direct feedback to the dataset stewards.

1.3. Value Proposition to Agencies

Agencies, via their Chief Information Officer (CIO) appointed Points of Contacts (POCs), are the key partners for populating Data.gov with high value, authoritative datasets. The POCs should establish relationships with the program offices which own/house data and determine the appropriate datasets for Data.gov. Since most agencies have information dissemination as part of their mission, Data.gov is a key component for improved mission delivery. In short, agencies that generate datasets are generally

tasked with sharing these datasets. Data.gov is a delivery channel to enable agencies to make their datasets more accessible, discoverable, and usable.

The Data.gov difference is both in the standardization of data discovery as well as the increased awareness and visibility of the site itself. As seen in Figure 2, Data.gov generates over 150,000 hits per day and drives this traffic to agency web sites in a measurable manner.

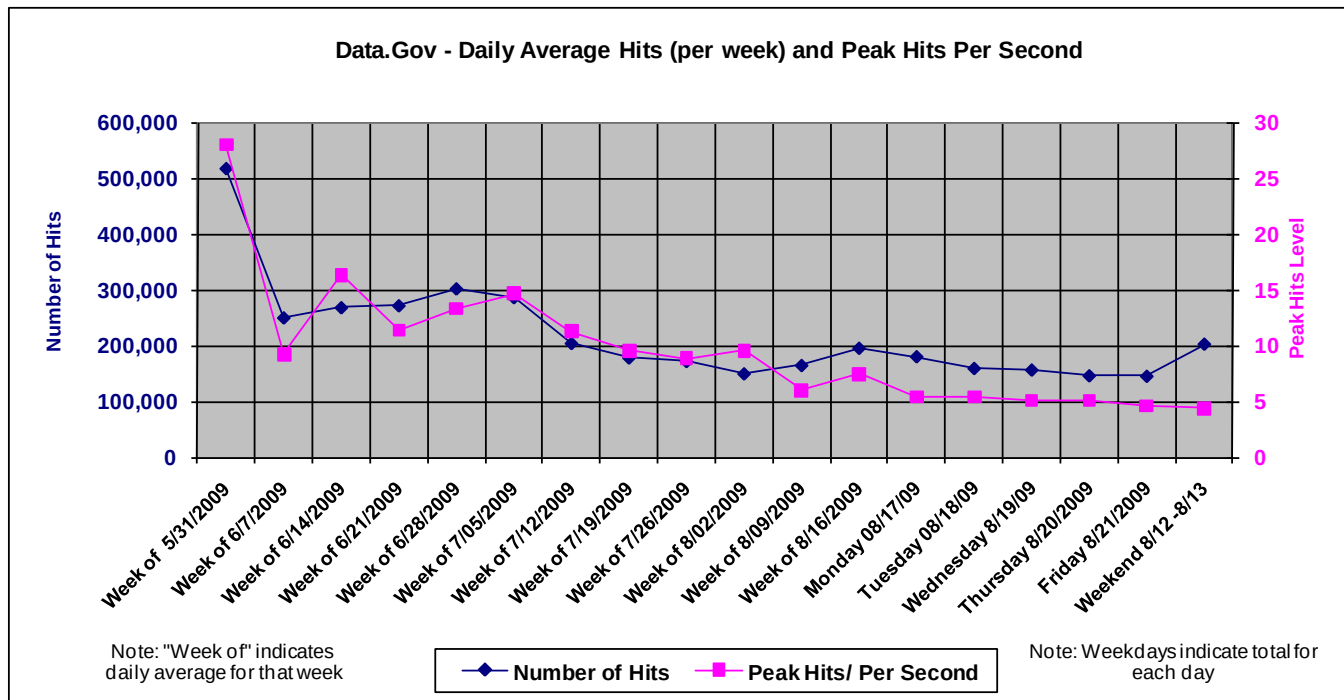


Figure 2: Daily Average Hits (per week) and Peak Hits Per Second (Data.gov)

The visibility and subsequent traffic on Data.gov are expected to increase with the addition of new agency datasets and tools, improvements in metadata, and improvements in data quality. By this logic, the benefit to participating agencies increases as additional agencies begin to participate more actively. In this manner, agencies have a vested interest in not only their own active participation, but the active participation of their peer agencies.

From an individual agency perspective, one value proposition of Data.gov is that it gives high visibility to datasets that they want to share with the public. As shown in Figure 3 and Figure 4, Data.gov includes five revolving panels of "featured tools and datasets" that provide even more visibility and showcase high-quality tools and datasets provided by the Federal government. These five featured tools and datasets are rotated on a regular basis to keep the content fresh and representative of the many topical areas within the purview of the Federal government. Data.gov is realizing the promise of an open and transparent government by steadily increasing and improving the public's access to the vast repository of information within the Federal government.

FEATURED TOOL:
U.S. GEOLOGICAL SURVEY (USGS)
 USGS Global Visualization Viewer for Aerial and Satellite Data

Ten million archive images of the Earth's surface are available for immediate selection and free download via the USGS Earth Resources Observation and Science (EROS) Center's Global Visualization Viewer. Users can preview thumbnails, browse images and download full-image selections from 1.5 million aerial photos of U.S. sites and 8.5 million images captured worldwide by U.S. Earth-observing satellites.

[VIEW THIS TOOL ▶](#)



Figure 3: DOI (USGS) Featured Tool Panel Example

FEATURED DATASET:
ENVIRONMENTAL PROTECTION AGENCY (EPA)
 EPA Toxic Release Inventory



The Toxics Release Inventory (TRI) is a publicly available EPA database that contains information on toxic chemical releases and waste management activities reported annually by certain industries as well as federal facilities. The TRI is available for each state for years 2005, 2006 and 2007.

View National data file of all US States and Territories:
 2005 [view ▶](#) 2006 [view ▶](#) 2007 [view ▶](#)

Figure 4: EPA Featured Dataset Panel Example

Data.gov assists agencies with their information dissemination requirements, and also provides agencies with a new and important public feedback mechanism. As Data.gov continues to evolve, agencies will be provided with new and more robust ways to obtain feedback directly from the end users of their authoritative datasets. Not only do agencies have a highly visible data dissemination channel, but agencies also have a streamlined and more direct way to receive feedback into their mission functions.

For instance, the public can provide specific narrative feedback on published datasets and tools. In these instances, agencies receive the benefit of direct feedback without the burden of investing in their own feedback mechanisms.

Datasets, tools, and query points that are published through Data.gov are required to have conforming metadata to promote discoverability and use. By participating in Data.gov, agencies receive the benefit of data stewards formulating the proper metadata and ensuring that the metadata are accurate for dissemination. In many cases, it is difficult to promote a standard approach to providing metadata within an organization unless they are being actively used. Since Data.gov uses the metadata on its site, there is a direct incentive for dataset stewards to ensure that their metadata are of the highest quality. The same is true for the quality of the data themselves. As datasets, tools, and query points become more transparent and more widely used, feedback from the public can be a market force for improving the quality of the data. Improved data quality not only benefits the public, but also supports improved policy-making and management within the Federal Government.

Agencies that actively participate in Data.gov are not only sharing their data, but are also increasing mission awareness. Active participation in Data.gov increases overall visibility and can enable a greater trust and appreciation for agency missions, their roles, and their overall performance in the service of the country. Transparency of agency datasets provides the public with the ability -- either through government tools, third-party web applications, or other means -- to understand their government and its impact on their lives.

Agencies also have another unique value proposition and that is to collaborate with other Federal organizations sharing a common data area. For example, several separate Federal organizations manage water data which creates an opportunity for these agencies to collaborate on posting, and describing their data in a way that increases access for the public and access for those agencies as well.

A final value-added proposition for agencies is that Data.gov acts as a living laboratory in which the problem of how to improve the public's ability to discover, access and reuse Federal Government data is actively being worked. Data.gov is a laboratory for new approaches to search, metadata, and visualizations and the knowledge gained is open to Federal agencies.

1.4. *Measuring Success*

Data.gov delivers value through measurably increased compliance and more consistent implementation of dissemination policies across the Federal Executive branch, a more consistent approach to dissemination and use of Federal Executive branch data, and increases in the span and scope of resulting economic, natural resource, and social applications. Data.gov serves a wide array of end users with value propositions that reach from research to knowledge discovery to fulfilling general intellectual curiosity. Measuring the success of Data.gov cannot be accomplished with a single metric like website page views or unique visitors. More appropriately, measuring success of Data.gov needs to include a variety of techniques that can capture both quantifiable and more qualitative signs of success.

The foremost quantifiable category of measurement for Data.gov is the actual use of the data assets. This category of measurement includes retail metrics such as page views, data queries, data downloads, and external "run time" access of hosted query points. This category of measurement also includes more wholesale metrics such as the count of agency and third party tools and mashups and the volume

and content of end user feedback. In addition to quantifiable metrics, Data.gov is also measured on more subjective assessment criteria such as the increased use of data-driven decisions.

Data.gov, working with agencies through the Points of Contact will establish a consistent performance framework for measuring the success and value creation from dissemination of Government information for use by both Data.gov and agencies.

2. Data.gov Operational Overview

2.1. Functionality of Data.gov

Data.gov is a three tiered web design model. The first tier features a home page with easy navigation to the catalogs, user feedback options, featured datasets and tools, and tutorials for assistance. The set of five featured datasets and tools are replaced periodically to promote the datasets of different missions throughout the government. The home page and its features are shown in Figure 5.

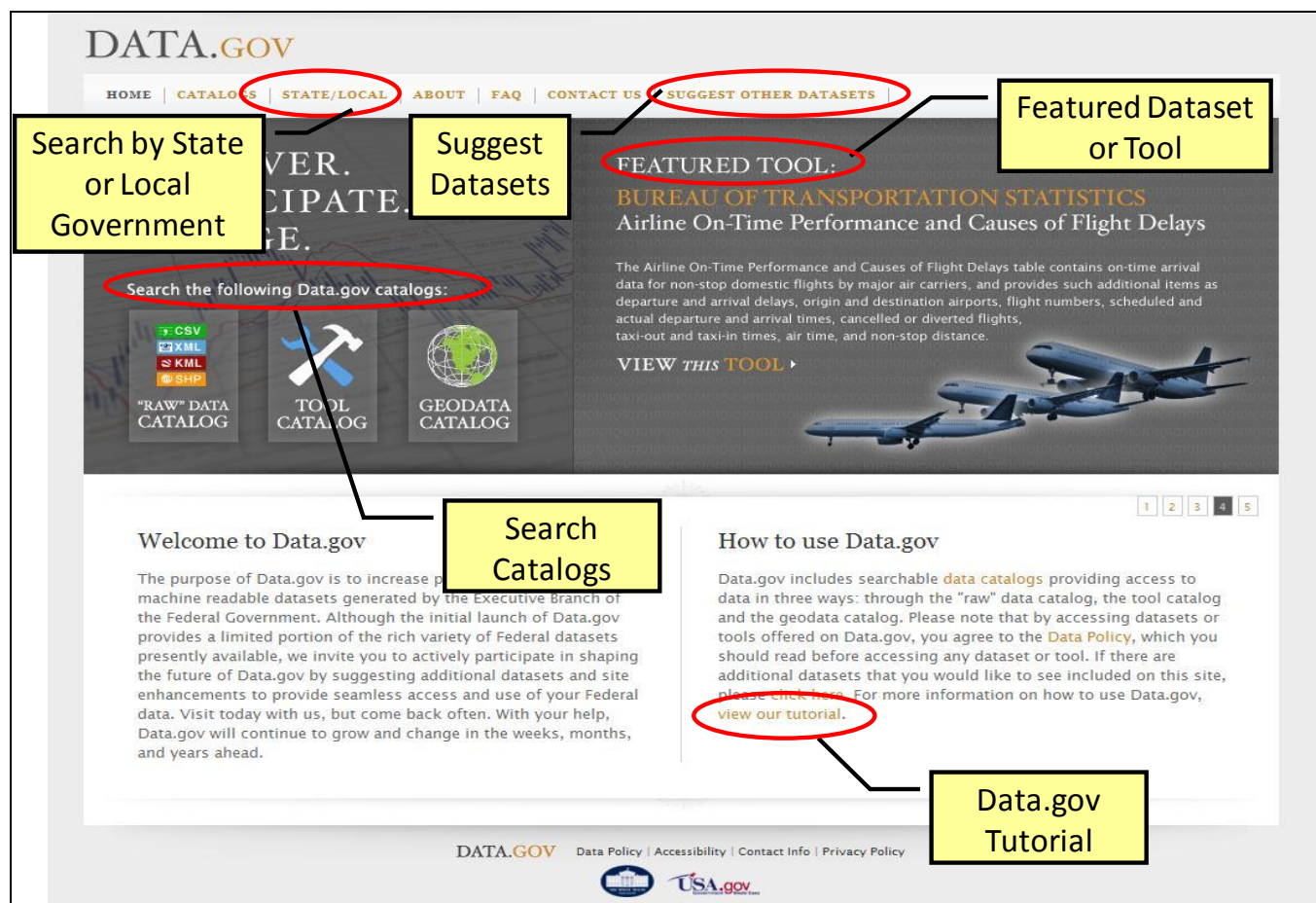


Figure 5: Data.gov Home Page

The second tier of the Data.gov design includes three catalogs:

1. "Raw" Data Catalog⁶: Features instant view/download of platform-independent, machine readable datasets 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 "raw" 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 datasets in the "raw" data catalog and the Geodata Catalog technically contains both "raw" data and tools.

The Tools Catalog and its features are shown in Figure 6.

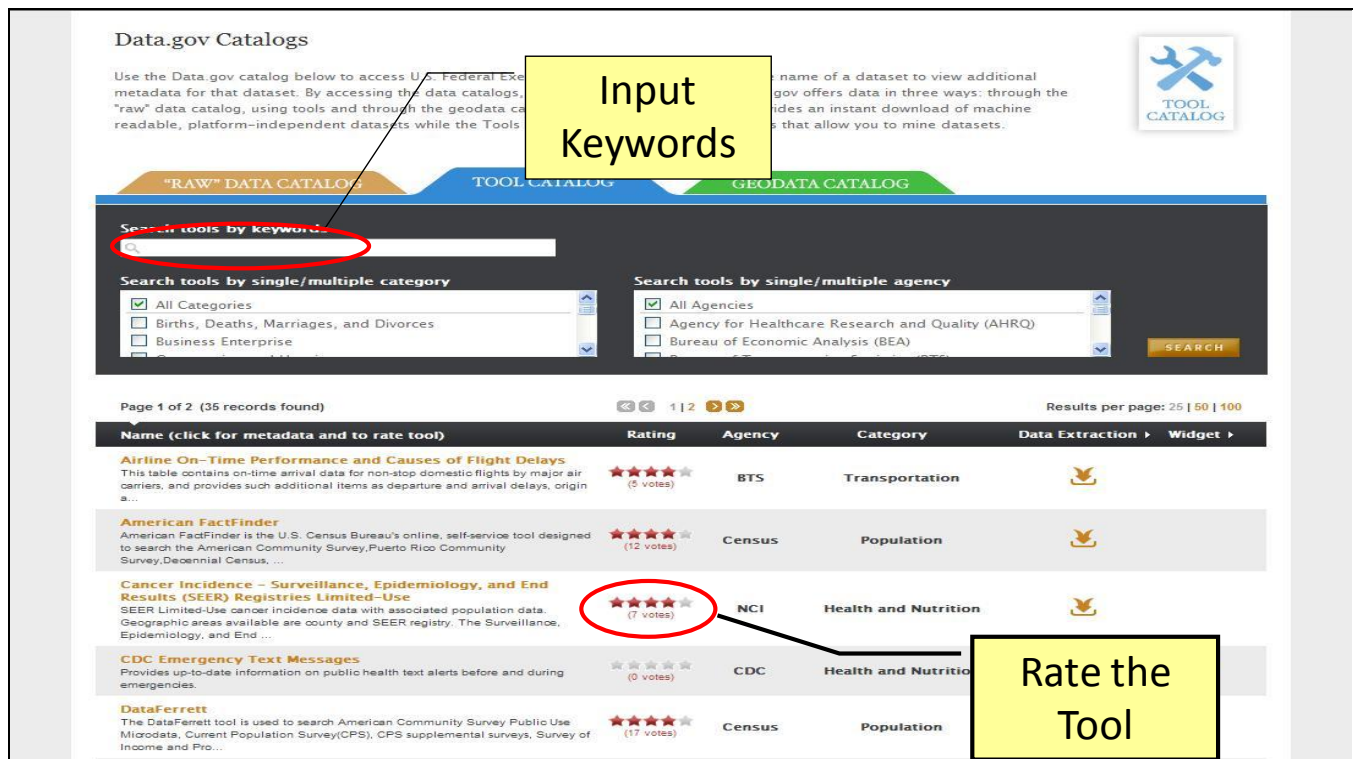


Figure 6: Data.gov Tools Catalog

The third tier of the Data.gov design 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. This third tier view and features are shown in Figure 7. The end user will access 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 will be able to download the datasets and/or tools, rate the data, and even comment on the dataset or tool.

⁶ 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.

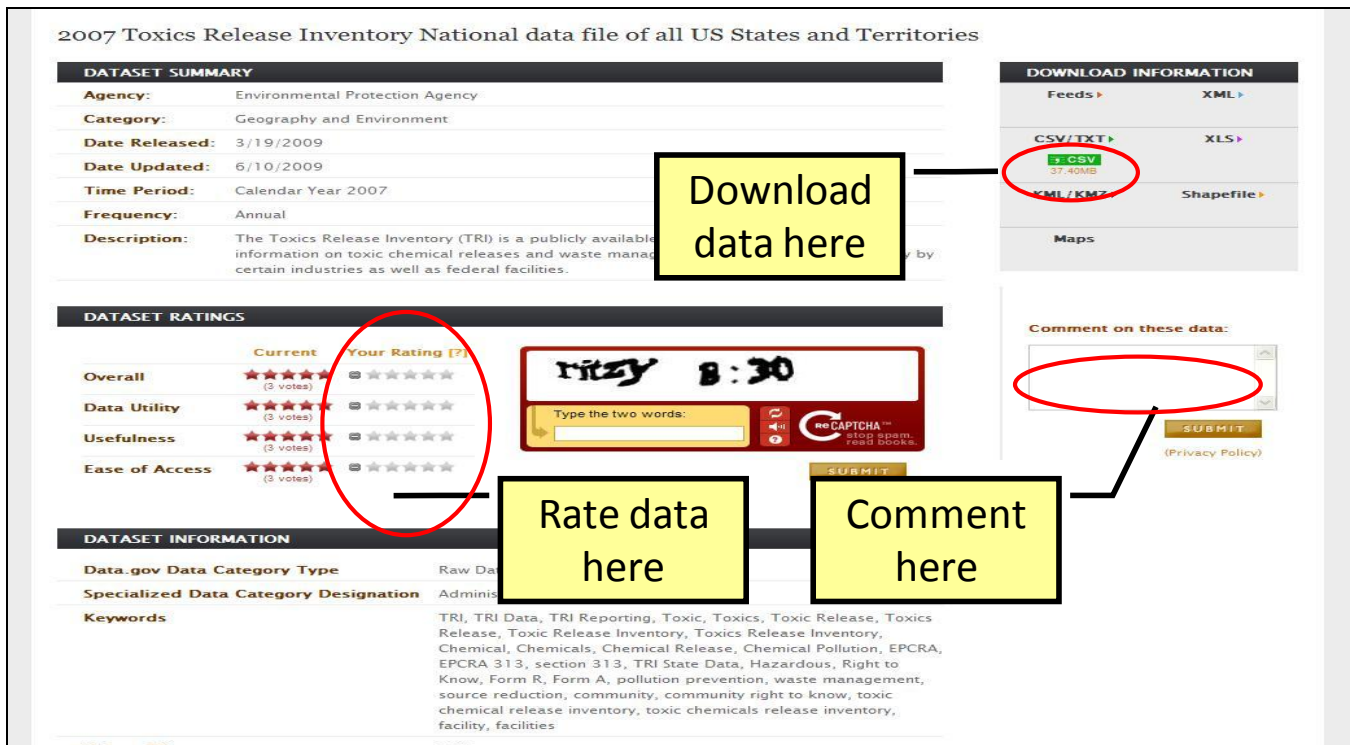


Figure 7: Data.gov Metadata Page

2.2. How Datasets and Tools Get to Data.gov

The Data.gov initiative was established by the Federal Chief Information Officers' Council and the E-Government and Information Technology Office at the Office of Management and Budget. 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 from 11 agencies.

Data.gov was launched on May 21, 2009, to serve as the single access point for publicly available authoritative Federal data. As part of the initial launch, a sequence of data calls went to Executive Branch agencies soliciting nominations for agency datasets that were already being disseminated to the public. The initial goal of Data.gov was to publish data characterized by minimal risk and high value to potential end users.

Since the initial launch, over 100,000 individual data resources are accessible through Data.gov. These data resources include raw datasets 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 raw datasets and tools. Table 1 details the August 25, 2009 view of datasets and tools on Data.gov from each participating agency/organization. Note that some "tools" counted in Table 1 actually represent hundred, 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. In the future, Data.gov will include more visible reporting of agency/organization participation.

Table 1: Datasets and Tools by Agency/Organization (as of 8/25/09)

Agency	Entries in Raw Data Catalog	Entries in Tool* Catalog	Entries in Geospatial Catalog	Total
Department of Agriculture	0	2	0	2
Department of Commerce	15	9	108,137	108,161
Department of Defense	0	122	0	122
Department of Education	0	2	0	2
Department of Energy	3	0	0	3
Environmental Protection Agency	296	3	174	473
General Services Administration	0	2	0	2
Department of Health and Human Services	60	25	0	85
Department of Homeland Security	42	2	0	44
Department of Housing and Urban Development	0	5	0	5
Department of Interior	31	4	1,140	1,175*
Department of Justice	6	4	0	10
Department of Labor	35	0	0	35
National Aeronautics and Space Administration	0	3	626	629
National Archives and Records	0	1	0	1
National Science Foundation	0	2	0	2
Nuclear Regulatory Commission	0	0	0	0
Office of Personnel Management	0	0	0	0
Small Business Administration	1	0	0	1
Social Security Administration	2	0	0	2
Department of State	0	2	0	2
Department of Transportation	0	6	0	6
Department of Treasury	64	1	0	65
Department of Veterans Affairs	0	0	0	0
Executive Office of the President	4	8	0	12
Other	20	5	0	25
Total	579	208	110,077	110,864

Note: * Some "tools" counted in Table 1 represent hundred, thousands, or millions of datasets however; they are represented and counted as one tool.

To facilitate the standardization of data exposed via Data.gov, the Data.gov team, in collaboration with the Office of Management and Budget (OMB), the CIO Council Data Architecture Subcommittee (DAS), and the CIO Council developed a metadata template designed to provide users with information about the utility of each data resource for any specific use.

The current process by which agencies and others submit datasets to Data.gov is described in Figure 8. This process is standardized for all agencies and requires manual processing and the coordinated manual review of the metadata template by both the agency point of contact (POC) and the Data.gov program management office (PMO). Note that the process bifurcates for geospatial data published to the Data.gov Geo-data Catalog.

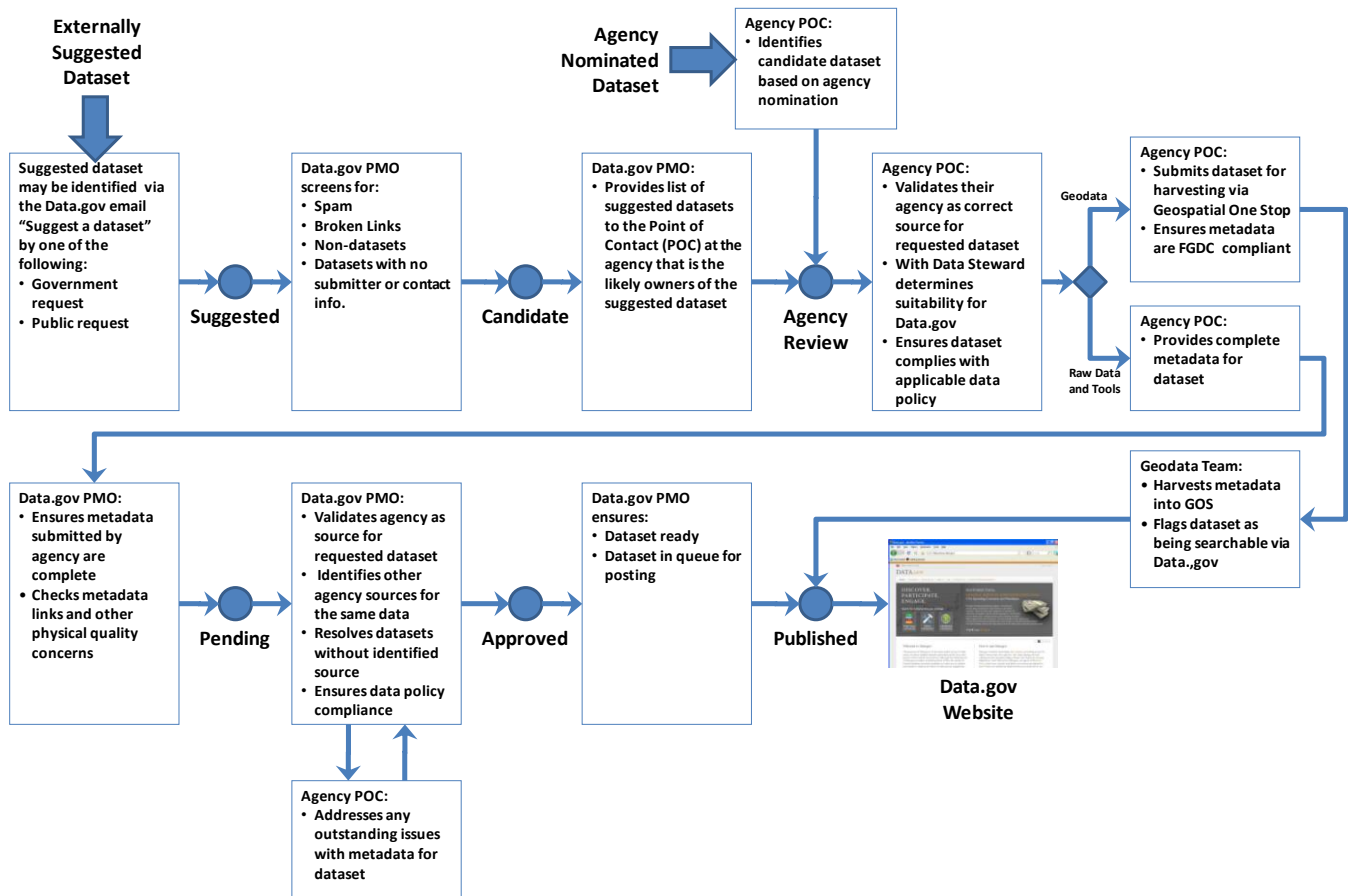


Figure 8: Current Process for Publishing Datasets in Data.gov

The agency POC plays a central role in completing the metadata template and ensuring that each submitted data set conforms to the [Data.gov data policy](#) (i.e., the agency is meeting all of its statutory requirements with respect to information quality, security, accessibility, privacy, and confidentiality). Once the metadata are completely specified, the dataset is submitted to the Data.gov Data Team which performs an additional check for completion. Once a dataset is complete, metadata is transmitted via e-mail wherein a final automated script is run on the metadata to check them one final time for completion and compatibility with the Data.gov database prior to publishing. Throughout these multiple quality checks, issues are communicated with the agency POC and resolved using e-mail. In order to ensure data quality standards are satisfied, agencies are required to certify that all data published on Data.gov are in compliance with agency data quality standards and other information quality assurance requirements. Presently there are 160 agency points of contact (POCs) working as a community to increase the datasets and tools available through Data.gov.

Metadata Template

Data.gov leverages a metadata template to document data elements or attributes about the specific content of data and data extraction tools, along with information about their use and access (where it is located, how it is associated, ownership, etc.). The Data.gov metadata template also includes additional 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. In developing the metadata template for Data.gov, metadata standards used in various communities of interest were reviewed and assessed. The core elements in the Data.gov metadata template were based on the *Dublin Core Metadata Element Set, Version 1.1* standard. The resulting metadata template now in use for Data.gov plays three roles:

- Enables users to identify what data they are searching for and where to find it
- Drives the population of the data catalogs within Data.gov
- Provides the platform for agencies to nominate datasets and/or tools for Data.gov

As such, the most critical elements of the metadata template include the data descriptions, keywords, data sources, and URLs of 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. 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”. We envision that communities of interest, both within the Executive Branch and outside, will recommend additional metadata elements for other domains.

For geospatial datasets and tools, the Data.gov metadata template only includes datasets and tools that are compliant with the Federal Geographic Data Committee (FGDC) Content Standard for Geospatial Metadata (FGDC-STD-001-1998). The FGDC content standard is the current requirement for datasets posted on the Geospatial One Stop (GOS) (geodata.gov). Because the goal of Data.gov is to provide access to all datasets 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 datasets 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.

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 datasets include additional metadata that describe relevant information about the datasets such as the spectral content, geospatial coordinates, image data quality (e.g., cloud cover), and other relevant attributes.

Geospatial Data Publishing Process

Agencies that have geospatial data that can be published to Geospatial One Stop (GOS) are encouraged to leverage the existing GOS processes for exposing their geospatial data assets. During the GOS submission process, agencies are permitted to select if they would like their geospatial data to be

simultaneously exposed through Data.gov. In these instances, the metadata associated with the geospatial datasets are harvested directly from the GOS catalog and exposed via the GOS application programming interface (API) to Data.gov. The Data.gov metadata are automatically populated based on pre-defined mappings between the GOS standard metadata and the Data.gov metadata template.

The automated use of the GOS API eliminates many of the manual quality checks performed by the Data.gov team. As with general datasets, the information quality, security, accessibility, privacy, and confidentiality aspects of the geospatial data and metadata remain an agency responsibility.

2.3. *Hosting Data.gov Using Cloud Services*

Data.gov is a premiere example of leveraging elastic cloud services to handle peak loads of traffic on a highly visible site. Just as USA.gov leverages cloud services, the Data.gov environment leverages cloud services including a content delivery network. The site traffic is largely handled through the content delivery network with very little impact on origin servers. Figure 9 summarizes the conceptual hosting environment.

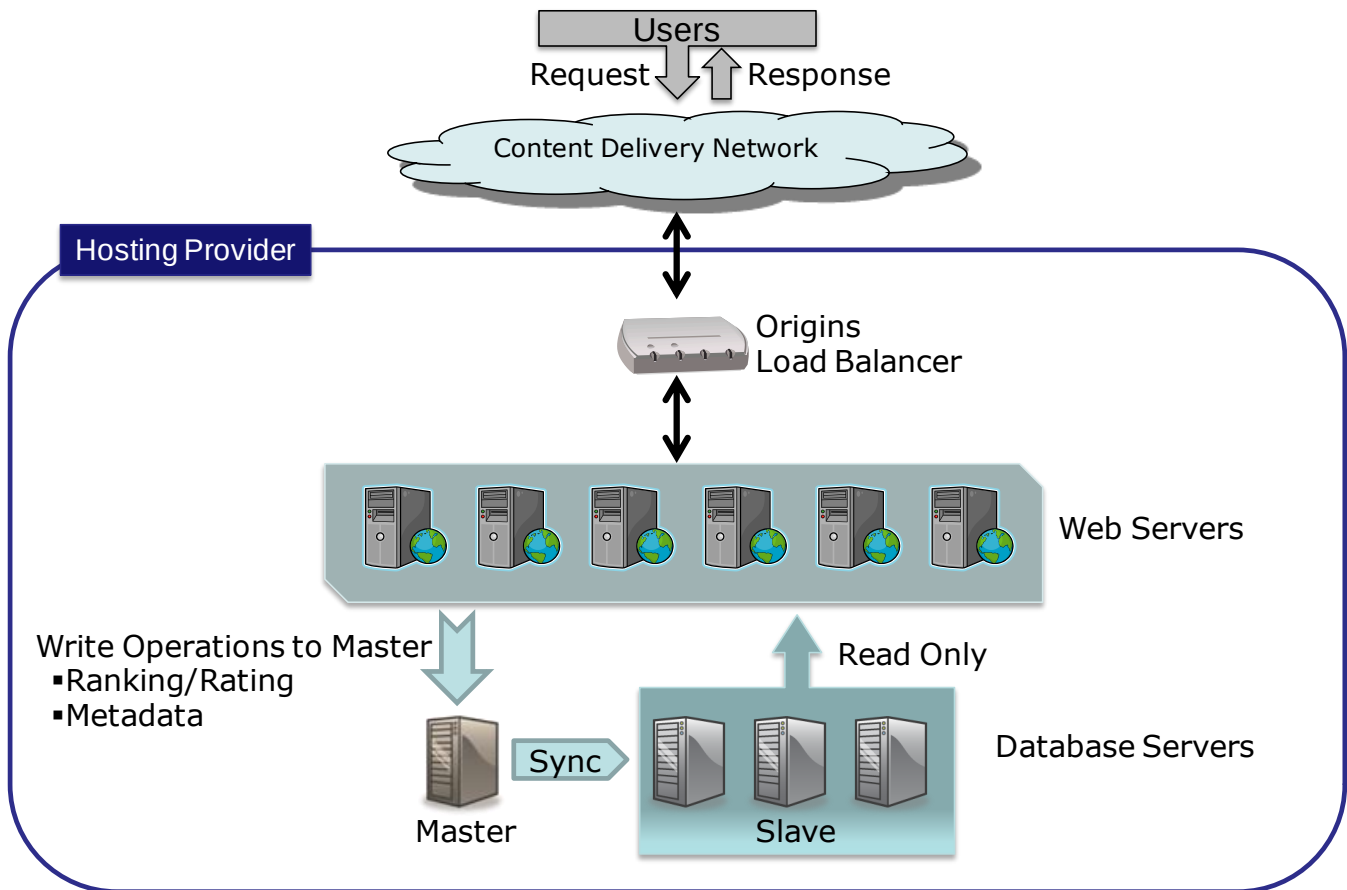


Figure 9: Conceptual Architecture Overview

2.4. **Who Makes Data.gov Work: Key Partnerships**

Agency Roles and Responsibilities

Agencies are the creators and managers of Federal datasets and related tools. The missions performed within the Federal Government generate large amounts of data that have the potential to be made available to the public (i.e., data whose distribution is not sensitive due to national security, privacy, or confidentiality considerations). Such data should be shared through Data.gov and the project management team is actively working with agency points of contact to increase the number of datasets that they post to Data.gov. As noted previously, maximizing the number of eligible datasets that are posted on Data.gov increases benefits to both the public and the programmatic missions of the agency by increasing the visibility and awareness of the data as well as providing direct feedback from end users who are leveraging the data.

Although most of the communications to date have been from OMB's E-gov/Federal CIO office, it is important to note that CIOs and their points of contact (see below) cannot populate and maintain Data.gov on their own. The program offices within agencies, which are the owners and stewards of the data assets that are being made available are key partners in Data.gov. OMB and Data.gov leadership has been and will continue to reach out directly to and engage with program offices.

Agencies' functional roles related to Data.gov:

- Agency program offices are the source of the data that are posted to Data.gov.
 - Program offices are responsible for determining which datasets and tools are suitable to be posted on Data.gov, being mindful of the significance of exposing datasets through Data.gov in terms of the implied authoritative and high quality nature of data associated with the Data.gov brand.
 - Program offices 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 mission. The term 'data steward' is used to refer to the agency staff that are directly responsible for managing a particular dataset.
- Agency program offices are responsible for ensuring that the data stewards for a particular data asset complete the required Data.gov metadata template for each dataset or tool to be posted on Data.gov.
- Agency program offices should facilitate Data.gov POCs and CIO efforts to understand and catalog data assets, as indicated below.
- Agency program offices, 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 the information quality, security, accessibility, privacy, and confidentiality.

- Agency CIOs, in conjunction with agency program partners, are responsible for cataloging⁷ and understanding their data assets, establishing authoritative sources, and ensuring 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 between mission and program performance and information resources to improve the efficiency of mission performance; and
 - 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 sector⁸.
- Agency CIOs and program offices 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 generally machine readable formats such as XML. In this role, agencies can determine if it is appropriate to make available a given data asset as a download, query point, or tool. Query points will increase in popularity as the general public application community requires easier access to data via web query.
- Agency CIOs have the responsibility for assigning an overall Data.gov "point of contact" for their agency (POC). The agency's Data.gov point of contact (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.

⁷ 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."

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

- The data steward is responsible for ensuring that the dataset is compliant with information quality guidelines and other applicable information dissemination requirements, that the corresponding metadata are compliant with the Data.gov template and are complete, and that the dataset is available online through the agency's website.
- The POC is responsible for understanding the Data.gov processes, the Data.gov metadata template, and the compliance requirements and for coordinating dataset submissions for the agency.
 - POCs are responsible for establishing an internal (agency) process for working with the program offices to ensure identification and evaluation of datasets 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 is also responsible for facilitating feedback to Data.gov from the data stewards regarding improvements to the metadata template, including recommendations generating taxonomies that facilitate interoperability.
 - The POCs are also the front line for agencies to respond to data requests and comments received from the public through Data.gov.
 - Critical success factors for POCs include:
 - Establishing, populating, and moving agency data sets and tools through a "pipeline" culminating in inclusion of the same in Data.gov.
 - As Data.gov matures, bring back to their agency, program office, and data stewards concrete feedback on the value generated and feedback garnered via Data.gov
 - Serving as a conduit for agency participation in the evolution and successful realization of target outcomes for Data.gov.

Senior Advisory Council

The Executive branch hosts many interagency efforts that focus on missions associated with data policy, information management, and enhancing the dissemination of digital data. Representatives of many of these formal communities of interest are being recruited to serve on a council of senior advisors to Data.gov. Their input will serve both to ensure alignment across the Executive branch with respect to modernizing information dissemination as well as motivating their colleagues in their home agencies to become more active in making their data assets available to Data.gov. Examples of these formal communities of interest that address data policy within the Executive Branch include:

- The Interagency Council on Statistical Policy (ICSP), which includes representation from the principal Federal statistical agencies;

- The Federal Geographic Data Committee (FGDC), an interagency 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 interagency working group of senior scientific and technical information (STI) managers from [13 Federal agencies](#);
- The Interagency 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.

Subject Matter Expert Technical Working Groups

Data.gov will harness the interests and expertise of staff across the Government when it stands up technical working groups designed to develop approaches to 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 that strengthen the interoperability of Federal datasets.

Federal Communities of Interest/Information Portals

Other interagency efforts, such as Science.gov and Fedstats.gov, are focused on serving distinct communities within the general public 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 such as those cited above. 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 adopt the Data.gov metadata template and are encouraged to expose corresponding data either as downloadable datasets, 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, and usefulness of the data.

State, Local, and Tribal Involvement

Although Data.gov does not catalog State, local, or tribal managed datasets, there is a shared benefit of cross-promoting efforts to catalog and make their data assets more transparent. State, local, and tribal governments are encouraged to leverage the thoughts, ideas, and execution patterns used by Data.gov to develop their own Data.gov style solutions. State, local, and tribal governments are also encouraged to inform the Data.gov team of their own implementations so that the Data.gov team can link to those specific sites from <http://www.data.gov/statedatasites>.

Furthermore, State, local, and tribal governments are encouraged to innovate new and interesting ways of cataloging, presenting, searching, and visualizing their datasets. State, local, and tribal governments are also encouraged to find more interactive and elegant ways of interacting directly with the public. As innovations are implemented, non-Federal governments are encouraged to share these breakthroughs with the Data.gov team for potential use on Data.gov.

2.5. *What Enables Data.gov: The Dataset Management System*

The Data.gov team will be launching a more automated approach for agencies to populate Data.gov. The Dataset Management System (DMS) will facilitate agencies' efforts to organize and maintain their Data.gov submissions. The Data.gov DMS will provide the necessary tools and capabilities to provide an automated manageable process for publishing datasets into the Data.gov catalog. The Data.gov PMO is constructing the DMS to be flexible enough to accommodate many distinct agency requirements, policies, and procedures.

Submitting a New Dataset

The Data.gov Dataset Management System allows the dataset submission originator to login by using their .gov or .mil email address and the password provided by their agency point of contact (POC).

DATA.gov @ GSA Home Dashboard User Management Logout

Welcome to Data.gov Dataset Management System (DMS)!

You may filter the result set by using the appropriate filters or add a new record.

Draft	Agency Review	PMO Review	QC Review	Published	TOTAL
29	62	30	30	288	439

Title and Description Filter By Status

- Past East Pacific Storm Tracks**
The National Weather Service (NWS) maintains historical Eastern Pacific Ocean hurricane weather data in a
PMO Review
- Tax Year 1996 SOI Exempt Organization Study**
Forms 990 and 990-EZ filed by IRC Section 501(c)(3)-(c)(9) organizations for Tax Year 1996
PMO Review
- Tax Year 2005 SOI Exempt Organization Study**
Forms 990 and 990-EZ filed by IRC Section 501(c)(3)-(c)(9) organizations for Tax Year 2005
PMO Review
- 2005-2007 American Community Survey Three-Year PUMS Housing File**
National survey that collects data from a sample of the resident population in the United States. Housing
PMO Review
- NOAA Watch Web Widgets**
This site contains widgets that track river levels and forecasts with content provided by NOAA. The widgets
PMO Review
- SESTAT**
National Survey of College Graduates (NSCG), National Survey of Recent College Graduates (NSRCG), Survey
PMO Review

[Add New Record](#)

Figure 10: Dataset Management System Home Page

If the originator does not have a password or has forgotten the password, they can click on a link on the DMS homepage which will email their agency POC. The POC passwords are logged within the Office of

Management and Budget's online collaboration system called MAX Community. Originators' passwords are managed by the agency and sub-agency points of contact.

The DMS allows the originators to submit new datasets and review the status of previously submitted datasets. 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. After an originator has submitted a dataset, an email notification is sent to both the originator, confirming their submission of the dataset, and the agency POC to inform them that a dataset is waiting for approval.

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 is built with the flexibility to allow agency POCs to edit, delete, approve, or return records to the originator. At any time the POCs have access to a dashboard that provides their agency's published, pending, approved, and not yet published datasets.

The Data.gov PMO can access the DMS to view a dashboard illustrating pending submissions, all published records to date, and all validated but not published records. The DMS will allow the PMO to validate records, return records to a POC, and edit, delete, or change published records. Although the DMS will allow for changes to be made to the records after they are published, typically these functions will be used by the PMO only in the case of an emergency.

Upon validation, the DMS will allow the PMO to select either 'publish' or 'publish immediately'. After a publishing selection has been made, the DMS sends an email notification to the agency POC and the submission originator informing them of the status. To ensure proper quality control, the PMO will have a final review of the data to validate its completeness and correctness within each field of the submitted record. Once final quality control is completed, the PMO is able to publish the dataset to Data.gov so the information will be available for public access. The PMO is currently exploring the use of feeds to eliminate some of the existing publication process. Feeds will allow agency POCs to have a more direct route to loading datasets and tools for inclusion on Data.gov.

3. The Path Forward

3.1. *Enhanced Requirements*

As users leverage Data.gov, feedback will be generated, user patterns will be discovered, and additional requirements will be identified. The Data.gov PMO is managing a process to evaluate the need to evolve the functional and technical requirements of Data.gov. Additional project documentation such as the Data.gov Project Plan and the Data.gov Functional and Technical Requirements document will capture the requirements and sequencing of planned implementations.

The current focus is on:

- streamlining the availability of datasets on Data.gov,
- supporting dataset metadata management at the source rather than in manual processes,
- improving agency control over data sets and tools presented on Data.gov,

- enhancing Data.gov visualization and feedback mechanisms to complement the raw datasets,
- strengthening the public feedback and response cycle is more visible and effective, and
- allowing agencies provide query points to directly use data rather than just making data downloadable.
- providing more useful, aligned and granular utilization and performance measures to support reporting on Data.gov value creation in general and for agencies, programs, and data stewards.

Data visualization services could be delivered through Data.gov to include analytics, graphics, charting, and other ways of using the datasets available through Data.gov. Additionally, dataset use will be enabled through Data.gov by providing data that are disseminated in a form that allows for external use including visualization.

Providing data in the right format is as critical as providing the data themselves. For instance, data could be provided using query points such as RESTful web services, web queries, or application programming interfaces. 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 are able to directly measure "run-time" use of their data as opposed to just recording instances of data downloads. Also, given agency control over the query point, agencies would be able to better support access to most current, correct versions of data resources as well as more clearly understand downstream use and value creation resulting from their data resources.

The evolution of Data.gov will also focus on ensuring that Data.gov includes data in platform-independent formats. The usefulness of the data to the end user communities is maximized when services, feeds, and linked open data are provided. This approach was noted in two recent forums:

From the "Open Government Brainstorming" Sessions:

- "Adopt the latest innovative technologies for disseminating data, including Atom data feeds".
- "Allow citizens to build their own applications on top of government online services; for example, using a 'Services Oriented Architecture' approach".

From the Recovery National Dialogue:

- "This data must be put up using the current open standards as Linked Open Data. There is already a rapidly growing Linked Open Data cloud, to which US recovery data will be a welcome addition. The priority should be on the provision of, as we say, 'Raw Data Now!'."

3.2. Modernizing Agency Data and Metadata Publishing

The most significant adjustment in the path forward will be a modernized approach to agency dataset /tool publishing for Data.gov. Two key drivers for modernizing this aspect of the solution are: 1) agencies need to be fully and immediately empowered to present their authoritative data sources

through Data.gov and 2) the process for populating Data.gov can be streamlined by leveraging a self-service publishing model that is less resource intensive and more immediately responsive.

Figure 11 highlights the future state conceptual approach by which agencies will expose datasets, tools, and query points to Data.gov, while still maintaining responsibility for their internal Data.gov submission process and statutory responsibilities for information quality, security, accessibility, privacy, and confidentiality. As in the current process, candidate datasets, tools, and query points may be identified by either government or public sources. In contrast to the existing process, we propose a modernized process that allows each agency to define and publish the information in the Data.gov metadata template immediately to Data.gov. The goal of this process is to empower agencies to populate and maintain the Data.gov catalog. Note that depending upon the type of dataset, tool, or query point, agencies may publish datasets directly to Data.gov or to an intermediate catalog (e.g., Geospatial One Stop) to be integrated with Data.gov.

This model features a “publish and subscribe” approach to obtaining the information required in the Data.gov metadata template whereby agencies will publish this information and Data.gov will subscribe to the publication in order to receive the information for inclusion on Data.gov. In other words, the authoritative source for the ‘fitness for use’ information contained in the current Data.gov metadata template will be published as agency feeds, thus allowing Data.gov to subscribe to the feeds in order to maintain immediately accurate metadata available on Data.gov.

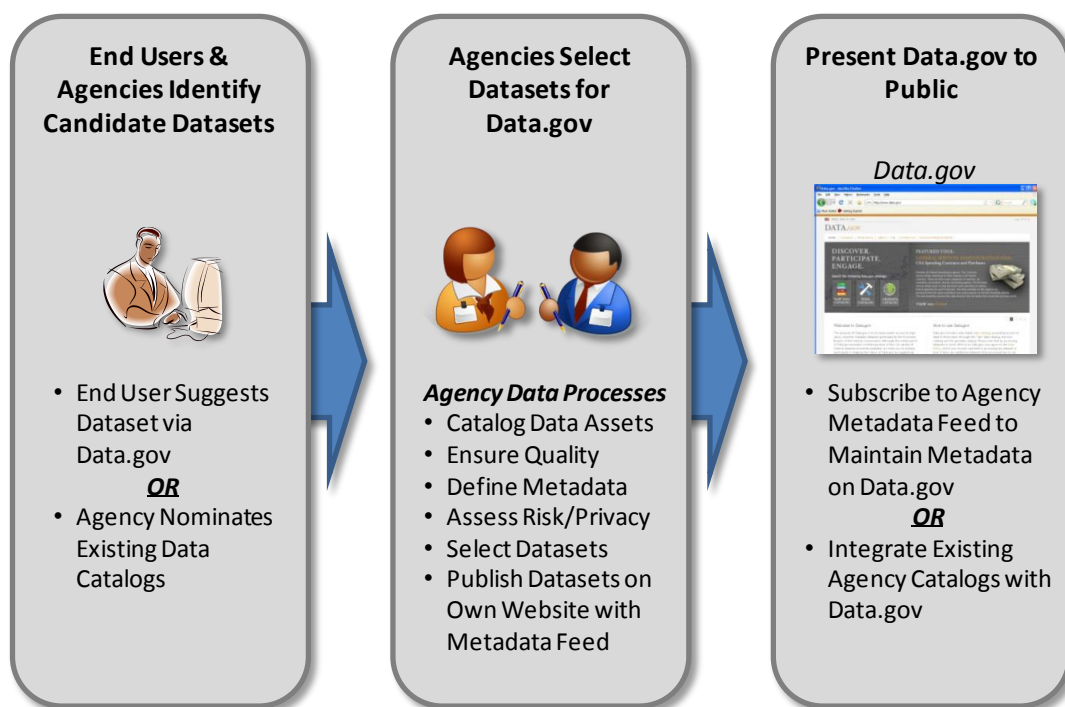


Figure 11: Data.gov Conceptual Agency Dataset Publishing Process

With the implementation of a more automated mechanism for maintaining the information necessary to populate the Data.gov metadata template, Data.gov will have the ability to perform automated scanning. This will be similar to the existing ability to scan FGDC metadata records currently employed within the GOS catalog. Data.gov will produce regular automated reports to agencies that identify

records with broken or invalid URL hyperlinks, records with incomplete or missing information, or records that otherwise do not comply with established policy. Agencies will be able to leverage this quality reporting to effectively monitor and control the process and identify and implement corrective actions as needed.

Figure 12 illustrates the high-level concept for the Data.gov publish and subscription process. This process is similar to existing established processes already successfully deployed for other catalogs including [Geospatial One Stop](#). Agencies are responsible for identifying and publishing metadata feeds that would be subscribed to by Data.gov and used to populate the Data.gov metadata repository. A key requirement of this model is that the agency business processes for publishing metadata feeds conform to applicable metadata template requirements. To supplement this approach, the metadata can also be provided using either the Data.gov metadata form (either a template or web form) or with the aid of a metadata tool that provides the Data.gov metadata template in XML format.

Important elements to consider with Figure 12 are:

- Agencies expose their metadata through feeds
- The feeds populate Data.gov and can be accessed by external websites, and search engines alike.
- Agency programs control which data assets are exposed and when.
- Data stewards control their own metadata directly
- The process is automated
- Federal information portals (e.g. science.gov) can also publish content to Data.gov

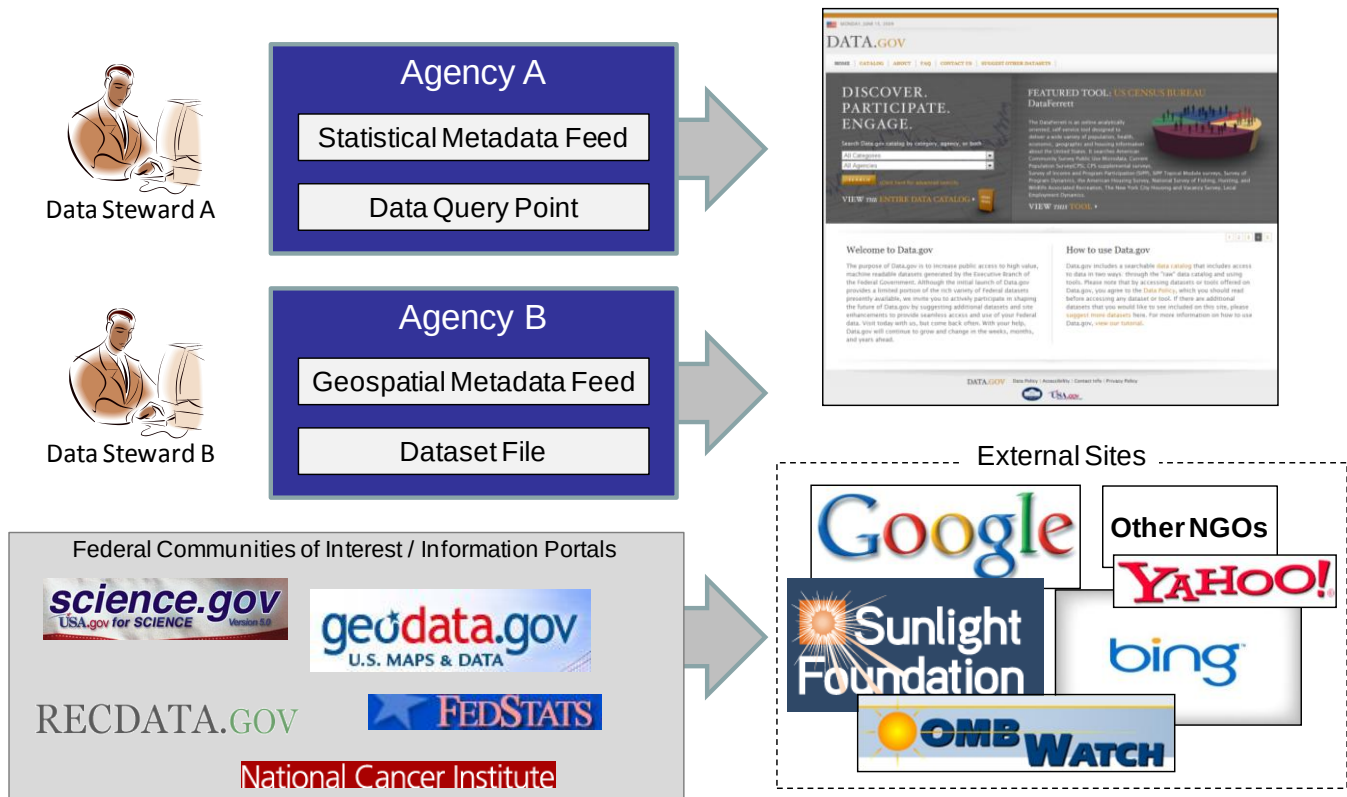


Figure 12: Data.gov Metadata Publishing Model
 (Representative external sites, no endorsement of external sites)

Note that datasets, tools, and query points all have completed Data.gov metadata information and the processes for accessing the feeds of the information required by the Data.gov metadata template are the same. Although the metadata model remains the same, the following long-term recommended best practices should be considered for agencies that are publishing tools:

- Agencies are encouraged to provide metadata for datasets served by data extraction tools, and, to the extent possible, migrate those datasets to the raw data catalog, assuming it can be done in a manner that protects privacy, confidentiality, and security.
- Where possible, agencies should include as part of any new investments the implementation of web-services as an integral component of the solution architecture for data extraction tools.

Data.gov, by design, instantiates the individual elements of the Data Reference Model v2.0 by separating the data, the metadata that describe the data, and the data query point. The separation of the three elements enables information sharing not just with access to individual datasets, but through an environment that facilitates the sharing and mash-up of data to enable the creation of new information, applications, and services. As such, Data.gov sets the standard by which all future public-facing Federal data can be architected and provides the standard, public-facing mechanism for publishing, sharing, and using such data.

3.3. *Assisting Agencies with Discovering Previously Published Data*

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. Across the many Federal websites, there exists the potential for agencies to have previously published data that have been lost within the complexities of today's domains, websites, and web-based directories.

The Data.gov team may provide the expertise 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. Note that the Data.gov team will not be scouring Federal websites and publishing the "found data" to Data.gov. The Data.gov team is considering deploying a search agent to scan Government domains to provide data to the agencies to assist in their data management practices.

Any deployed search agent would be officially sanctioned, comply with appropriate security specifications, and be approved in advance of use. The search agent would scan through Federal domains and formulate an index of potential datasets and build reports to deliver to agencies. This search agent and 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. This is intended not to automatically populate Data.gov, but rather to assist agencies with their own data inventory, management, and publication processes.

3.4. *Working with Other Federal Websites*

In general, Data.gov will be the source location to access raw data behind some of the government's most significant websites. Existing and newer websites such as USA.gov, the [Federal IT Dashboard](http://FederalITDashboard.gov), USASearch.gov, FBO.gov, USAspending.gov, and Grants.gov all have major presentations of data using search and presentations technologies. The raw data behind these websites will be part of the inventorying and metadata feed process as previously described. Many of the websites currently have web service mechanisms for accessing their raw data. Moving forward, services will continue to be implemented and offered and those services will be cataloged and presented through Data.gov

3.5. *Incorporating the Semantic Web*

[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 way in which "formal information models" are designed and implemented has changed significantly since the publication of [OMB Memorandum M-06-02](#); however the underpinning intent of structured description of data has not.

[OMB Circular A-130](#)⁹ states that "the application of up-to-date information technology presents opportunities to promote fundamental changes in agency structures, work processes, and ways of

⁹ OMB Circular A-130 states "(d) In determining whether and how to disseminate information to the public, agencies will:

interacting with the public that improve the effectiveness and efficiency of Federal agencies”. Furthermore, [OMB Memorandum M-06-02](#) further clarifies the existing obligation¹⁰ that agencies “provide on your principal websites search capabilities appropriately equivalent to industry best practices to assist the public in locating government information intended for public use”.

The bottom line is that today’s “industry best practices” are not grounded in basic search capabilities but are more frequently grounded in semantic techniques that enable the semantic web and query points that the public can directly access. 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 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. Semantically enabled techniques could include approaches such as [linked open data](#), [Information Exchange Package Documentation \(IEPD\)](#) approach associated with the [National Information Exchange Model](#), and [W3C standards such as RDF](#).

Data.gov can leverage semantically enabled techniques to allow traditionally disparate datasets to be connected via the semantic web. Agencies that expose datasets that leverage semantically enabled techniques are allowing end users to query those data in combination with other data that might be outside of the agency’s purview. Semantically enabled techniques then allow agencies to publish their own datasets that can be linked together not by physically storing and relating the data in one place, but rather by leveraging the semantically enabled techniques and web query capabilities to allow for queries that relate datasets from across the web. The semantically enabled approach eliminates the need for massive centralization of Federal datasets while making queries, visualization, and data discovery more powerful from within Data.gov and any other external data sites that leverage these Federal datasets.

Semantically enabled data is characterized as being open, modular, and scalable. The openness of semantically enabled data refers to datasets that are defined using open source or non-proprietary markup so that the data themselves can be accessed by any types of systems or applications.

-
- (i) Disseminate information in a manner that achieves the best balance between the goals of maximizing the usefulness of the information and minimizing the cost to the government and the public;
 - (ii) Disseminate information dissemination products on equitable and timely terms;
 - (iii) Take advantage of all dissemination channels, Federal and non-Federal, including State and local governments, libraries and private sector entities, in discharging agency information dissemination responsibilities;
 - (iv) Help the public locate government information maintained by or for the agency.”

¹⁰ [OMB Memorandum M-05-04](#), “Policies for Federal Agency Public Websites,”

The modularity of semantically enabled data refers to the ability to combine or mash-up semantically enabled data with other semantically enabled data. Essentially, the modularity of semantically enabled data allows the data to be used more easily in conjunction with other semantically enabled data.

The scalability of semantically enabled data refers to the ability to incrementally add additional semantically enabled data to existing data sets that are already using a semantically enabled data markup. For instance, you can add additional attributes or relationships to existing semantically enabled data using markup.

As semantically enabled data concepts are matured and pilots are successfully executed, the Data.gov team will provide specific guidance to help agencies implement semantic markup within their datasets. Specifically, the Data.gov team will not only leverage semantically enabled techniques within the Data.gov site, but will help agencies implement semantically enabled data within their datasets so that the datasets can be better leveraged not only by Data.gov but also by other end users of the data.

4. Transitioning to the Future

4.1. *Step 1: Establish Enhanced Governance*

As the Data.gov initiative has progressed, the governance has also matured and adjusted to the needs of the initiative. The path forward will include a continuing improvement to the governance structure including processes, tools, business rules, and team communications.

Establishing an enhanced governance structure is a key first step in the transition to the future of Data.gov. This important first step is being developed and documented by the Data.gov PMO and is being published in a separate Data.gov governance document.

4.2. *Step 2: Reinforce Existing Requirements and Issue New Standard Guidance to Agencies*

Many statutory responsibilities are applicable to Data.gov and will be reinforced in communications with agencies. Though additional guidance will be issued, existing legislative requirements are highly applicable to Data.gov.

Information Quality

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 to help agencies ensure and maximize the quality, utility, objectivity, and integrity of the information that they disseminate (meaning to share with, or give access to the public). 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.

The fact that Data.gov enables agencies to easily share information with a wide audience not only offers great benefits to society, but also increases the potential harm that can result from the dissemination of information that does not meet basic information quality guidelines. As a consequence, Data.gov

requires that agencies ensure that data disseminated via Data.gov are consistent with their agency's information quality guidelines.

Security and Privacy

Data.gov maintains the President's commitment to protecting the privacy of the American people. Therefore, all applicable privacy protections are enforced when the public interact with the Data.gov environment. Data.gov requires that data shared by agencies conform to all applicable security and privacy requirements including the [Privacy Act of 1974](#), the [E-Government Act of 2002](#), applicable Federal security standards including NIST 800-39, and other guidance as issued by OMB (See: [Office of Information and Regulatory Affairs \(OIRA\) Information Policy](#)).

Paperwork Reduction Act

The [Paperwork Reduction Act \(PRA\)](#) applied to Data.gov provides functionality by which the public is solicited to provide ratings of datasets. Prior to Data.gov releasing this functionality, OMB review and approval was obtained based upon an evaluation of the need to collect this information, the practical utility of the information, and the minimal burden imposed on the public in responding to the requested ratings.

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. In addition, all agency websites hosting data and data extraction tools that are exposed in Data.gov are also required to conform to applicable Section 508 accessibility requirements.

Additional Guidance

The [OMB e-Government Office](#) is the executive sponsor of Data.gov. OMB will issue guidance to agencies on how to participate in Data.gov and fulfill expectations for transparency, participation, and collaboration, in alignment with the [Open Government Memorandum](#) and [OMB Office of Information and Regulatory Affairs policies on information dissemination](#).

Guidance from the OMB will include an overview of agency responsibilities, a detailed outline of the Data.gov metadata expectations, and guidance for publishing datasets, geospatial data, and data related tools.

The guidance may address agency responsibilities such as ensuring that all datasets submitted to Data.gov by their respective agencies represent an authoritative source and meet the requirements outlined in legislation and OMB directives. The guidance may address agency specific metadata responsibilities such as maintaining their metadata in up to date formats and establishing an agency specific process for selecting and submitting metadata that conform to the agency process.

The guidance may include a detailed listing of the core and extension attributes for the information to be included in the Data.gov metadata template. Core metadata elements are required for all datasets

presented through Data.gov while certain datasets can also include special extension attributes that are specific to the type of dataset itself.

Lastly, the guidance may include detailed steps on exposing datasets, geospatial data, and data related tools to Data.gov. The processes for these data asset types are similar with small variations. The process for exposing data assets to Data.gov is evolving and the new, more streamlined process will be documented in the guidance as well as future releases of this Concept of Operations document.

4.3. *Step 3: Sequence the Implementation of the Enhancements*

The path forward includes the documentation and evaluation of future requirements per the processes and structure of the enhanced governance. These processes will be outlined in a series of living requirements documents that continues to change and evolve based on feedback and the needs of the larger end user community. The Data.gov PMO is managing a process to evaluate and determine formal functional and technical requirements for the evolution of Data.gov. These requirements are being evaluated and priorities are being assigned as well as the overall sequencing of releases. The Data.gov Project Plan is the official documentation for the sequencing of the implementation of the enhancements with clear scope, cost, and risk assessments for continued implementation success.

5. Reference

In addition to the references explicitly cited throughout this document, more information can be obtained by visiting the Data.gov Frequently Asked Questions: <http://www.data.gov/faq>.

Also you may refer to the Data.gov Glossary of Terms: <http://www.data.gov/glossary>