United States
Department of Transportation

Data.gov Working Group

Data.gov Interim Identification & Prioritization Process and Guidelines v1.0

June 2010

Revision History

Version	Overview	Author	Date
0.1	Initial Draft	Data.gov team	05/05/2010
0.2	Revised and expanded discussion in Section 3	Data.gov team	05/27/2010
1.0	Incorporated comments from the Executive Sponsor	Data.gov team	06/01/2010

Table of Contents

Section 1. Overview

What is Data.gov?

Data.gov is a component of President Obama's Open Government Initiative as expressed in Presidential Memorandum "Transparency and Open Government," dated January 21, 2009. The guiding principle of Data.gov is that Federal data and information should be "disclosed rapidly in forms that the public can readily find and use." Data.gov was developed and is managed by the Federal Chief Information Officer (CIO) Council.

Data.gov is explicitly concerned with open data (that is, data made available to the general public without restriction) that are generated, held, and formally published by the Executive Branch of the Federal Government. It is hoped that improving the accessibility of Federal data will foster innovation, lead to new discoveries, expand public-private communities of interest, and help fuel the knowledge economy.

The Data.gov System

Data.gov is a Web-based catalog system that assists the public with easily finding, learning about, downloading, and using datasets, associated applications, and services that leverage data. It relies on a simplified set of metadata and presents information in the form of catalog citations. For Department of Transportation (DOT) data, the Data.gov metadata is derived from the much more complex, specialized metadata that are used for internal and Government-wide scientific and geospatial cataloging, and by database systems.

Data.gov contains only citations describing datasets, data access tools, and services, and provides links to the original data sources. Data.gov does not store or serve copies of these assets. Originating agencies remain responsible for storing and maintaining access to them.

The Data.gov Dataset Management System (DMS)

The Dataset Management System (DMS) is an automated process for publishing datasets into the Data.gov catalog. It facilitates agency efforts to organize and maintain Data.gov submissions. The DMS enables Data.gov to generate its metadata catalog, which includes pointers to the agency source systems where the data is actually stored. In this manner, Data.gov provides a searchable inventory of data without housing data.

The DOT and Data.gov

Access to DOT data provides opportunities for consumers to conduct valuable research and analyses, combine data layers into new and interesting "mashups" of DOT and non-DOT data, and build novel applications, services, or derivative information products. Increased visibility and use of DOT data will result in increased citation, innovation and new research ideas. It will also lend greater credibility to scientific, engineering, and policy-making communities across a broad spectrum of the public and private sectors.

Section 2 provides information about the structure and content of Data.gov as they relate to the DOT. Section 3 outlines an interim publishing process and provides guidelines and resource links to assist DOT program managers and data access coordinators to identify, evaluate,

prioritize, and prepare datasets and tools for inclusion in Data.gov. The evaluation questionnaire is contained in Appendix A.

Section 2. What should the DOT publish via Data.gov?

Data.gov Organization and Content Framework

Data.gov is organized into separate catalogs to provide access to three classes of published, publicly accessible Federal information assets – raw data, geodata, and tools – each with specific attributes, as summarized in table 1.

Catalog	Defining Characteristics
Raw Data	Referenced datasets are considered nongeospatial if they are primarily organized and distinguished on the basis of subject or purpose rather than geographic location. Data are delivered in a variety of machine-readable, platform-independent forms for immediate use by consumer applications and do not need a DOT database system or interface for action.
Geodata	Referenced datasets are considered geospatial if they are primarily organized and distinguished on the basis of geographic location rather than subject or purpose. Data are delivered in a variety of machine-readable, platform-independent forms for immediate use by consumers. Datasets can be downloaded immediately without first passing through a local DOT database system or interface. The geodata catalog provides a direct connection to a subset of Federal geospatial data cataloged in the Geospatial One Stop (GOS) system.
Tool	Many data assets are only available through data access tools or applications. Examples include (1) datasets that can only be accessed in whole or part indirectly through an existing local DOT application-driven interface or portal; and (2) links to DOT data extraction, visualization and delivery applications and services, including Geographical Information Systems (GIS), application program interfaces (APIs), and tools through which datasets may be mined and portions extracted by consumers.

Table 1. Data.gov catalog content characteristics.

Working definitions of terms used throughout this document will ensure uniformity of the application these guidelines across all DOT Operating Administrations. These definitions are adapted from the Data.gov Dataset Management System User Guide.

- **Data:** a value or set of values representing a specific concept or concepts. Data become information when analyzed to extract meaning and to provide context. The meaning of data can vary, depending on its context.

- **Dataset:** an organized collection of data. The most basic representation of a dataset is the compilation of data elements presented in tabular form. Each column represents a particular variable. Each row corresponds to a given value of that column's variable. A dataset may also present information in a variety of nontabular formats, such as an extended markup language (XML) file, a geospatial data file, or an image file, and so forth. The Dataset Tool Catalog contains simple, application-driven access to Federal data, and the Geodata Catalog contains Federal geospatial data.
- **Metadata:** Describes a number of characteristics, or attributes, of data (that is, data that describes data, as defined by ISO/IEC 11179-3 or the Federal Geographic Data Committee (FGDC)). For any particular datum, the metadata may describe how the datum is represented, and what the ranges are for acceptable values. The datum should be labeled, and its relationship to other data should be defined. Metadata also may provide other relevant information, such as the responsible access coordinators, associated laws and regulations, and access management policy. The metadata for structured data objects describes the structure, data elements, interrelationships, and other characteristics of information, including its creation, disposition, access and handling controls, formats, content, and context, as well as related audit trails.

Data.gov – Dataset Formats

While XML is the preferred open format for Data.gov and DOT datasets, it is not the sole format for providing data. Acceptable formats (as defined in the Data.gov Dataset Management System User Guide) are summarized in Table 2.

Catalog	Defining Characteristics
Raw Data	XML, RDF, CSV, TXT, KML, KMZ, XLS, XLSB, ESRI Shapefile, ATOM, RSS, CAP, or other structured machine-readable form. Data embedded in HTML pages and PDF files (for example, in displayed tables) should be reviewed and traced to their source DOT information system. If data can be traced to a source system, it should be considered for publication in the Raw Data Catalog.

Catalog	Defining Characteristics
Tool	If single or multiple datasets are offered for download indirectly, they should be classified as one of the following tool types based on delivery mode: Data Mining/Extraction Tool, Data Feed, or Widget/Gadget. A Data Mining/Extraction Tool may be a database access facility, Web mapping, or data visualization application. It may also be a Web page (containing links to downloadable data files). Data Extraction Tool and Web page datasets are typically delivered using file compression formats such as ZIP, GZIP, and TAR. Feeds will be in XML formats, including ATOM, RSS, and CAP. Data Feed is….. Widgets/Gadgets are documented, shareable APIs and portable, standalone, embeddable data-access applets. Note: tools that require logins, explicitly restrict the data to less-than-full public use, or are otherwise incompatible with the Data.gov open data-access policy are currently not considered for publishing in the Data.gov Tools Catalog.

Table 2. Data.gov acceptable formats

Section 3. How should DOT prioritize data for publication?

Selection Criteria for Data.gov

All datasets shall be evaluated and scored against a uniform set of criteria. These criteria address value, quality, and manageability. Taken together, these criteria provide the basis for selecting and prioritizing the DOT's inventory of candidate datasets and tools. Each dimension is weighted as follows:

- Value: 40%
- Quality: 35%
- Manageability: 25%

Utilizing these dimensions, DOT will prioritize the release of data. The ideal dataset has a high value, high quality, and is easily managed. Once the inventory is complete, the datasets can be plotted and evaluated. A notional plot is provided here:

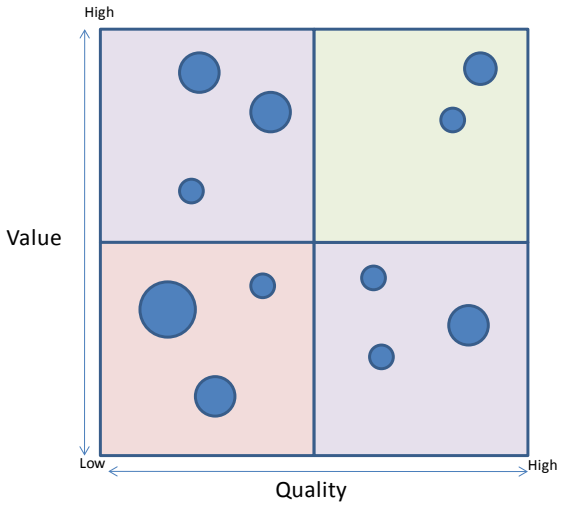

Figure 1. Illustrative Plot of Prioritized Inventory

Datasets that appear in the upper-right quadrant are high-value, high-quality datasets. Those appearing in the upper-left quadrant would be high-value datasets that require further analysis prior to release, and would be the first priority in addressing the release of additional data. Lower-value, high-quality data is found in the lower-right quadrant. These datasets are third priority on DOT's release schedule.

The DOT's inventory and scoring methodology will be fully automated through the DOT Services/Data Architecture Group (S/DAG) Metadata Registry at http://data.nhtsa.gov/.

Selection Criteria: Value

A dataset is considered valuable if it is relevant (internally or externally or both), usable, and readily available. To assess a dataset's relevance, the questionnaire in Appendix A evaluates its alignment to DOT's strategic priorities, its applicability to multiple modes of transportation (especially when supporting cross-modal analysis or decision-making), its usefulness to

application developers, its value in improving accountability, and the scope of its coverage – whether it can be useful at an individual, state, local, and/or federal level.

A usable dataset is fully defined and described, and is ready for consumption by an interested party. A fully defined and described dataset is one with a complete set of metadata thatexplains the data elements and permissible values. A dataset that is ready for consumption means that it is easy to understand without additional context. This section's questions evaluate whether the dataset can be understood by a consumer of that data, knowing that DOT's dataset users represent a ride range of consumers—including researchers, market analysts, policymakers, and the general public.

Finally, an available dataset may be one that is frequently requested—either through public consultation, through the normal reporting process (Congressional or annual reports), or through Freedom of Information Act (FOIA) requests. In addition, an available dataset is one that is already released on DOT Web sites, where the demand for that data can be monitored.

Selection Criteria: Quality

Quality is generally a subjective characterization, but it reflects the consumer's perception of its usefulness and reliability. Because data quality can be measured, and thresholds of acceptability can be established, the questionnaire integrates these kinds of measures in evaluating a dataset for release. The data quality section of the questionnaire evaluates a dataset's completeness and its consistency.

A complete dataset is one that is assigned a business owner and, if applicable, a data quality steward. The dataset is complete (has all of the required records) when data, especially those elements critical to analysis, are filled consistently. Derived values are clearly documented.

A consistent dataset is one that uses DOT, industry, or other appropriate standards to describe common data elements. The consistency of these master values in the data should permit the consumer to generate meaningful groupings within the data. For example, suppose a dataset reports contract information–the contract number should either always include or always exclude hyphens and spaces so that the field values can be aggregated appropriately. In addition, the value of a given data element has the same meaning, regardless of the context in which it is presented–for example, Vehicle Mile Travelled (VMT) should mean the same thing in every published dataset. If that is a derived field and formulas have been changed from release-to-release, such changed are documented so that the data are consistent over time.

Selection Criteria: Manageability

Manageability of release is evaluated on dataset sensitivity and the level of effort required to make the dataset releasable.

Sensitivity of a given dataset does not, in and of itself, preclude the release of a DOT dataset. Sensitivity is an independent variable in this scoring methodology. A sensitive dataset may still be high value and high quality, and there may be methods that can be used to suppress or generalize sensitive information to support releasability. Sensitivity includes privacy,

confidentiality, national security, or other non-public information relating to internal government operations. There may also be statutory prohibitions governing the release of a dataset. The characteristics are recorded during the inventory evaluation process to ensure that all sensitivity risks are known.

Further manageability considerations surround the level of effort required to make a dataset releasable. This question is designed to address a variety of actions that may be required to get a dataset in a releasable state. These include, but are not limited to the level effort required to:

- Convert data that is not into an open format into one that is open;
- Convert data that is a less usable open format into one that is more appropriate and more useful;
- Scrub sensitive information from a dataset so that it may be releasable;
- Convert the underlying data in a data mining tool into a data feed;
- Improve the quality of a dataset so that it meets minimum release thresholds;
- Improve the definition of the dataset so as to ensure it is fully defined & described; and
- Automate the extraction, transformation, and loading of a dataset into a repository on a regular release schedule, if applicable.

Appendix A. Data Scoring Criteria

Dataset Name	<Name of the dataset>
Dataset Format	<e.g. csv,xls,xml>
Dataset Source Business System	<if it is an application>
Owning Operating Administration (OA)	<Name of OA>
Owning OA Office (Routing Symbol)	<Routing Symbol>
Data Steward	<Name and Email>
Business Owner	<Name and Email>
Release Approval Authority	<Name and Email>

	Dimensions			Yes	No
40 = Highest Value	Relevance & Accountability	1	Does the data support the Secretary of DOT's current strategic priorities? *(That is, improve safety, protect the environment, support national security preparedness and response, reduce congestion for all Americans and increase global transportation connectivity in support of Nation's economy.)*	2.5	0
		2	Does the dataset have the potential to support multi-modal orcross-modal analytics and decision-making? *(that is, does the data support the mission of more than one operating administration)*	2	0
		3	Could the data enhance potential contribution to the creation of novel and useful third-party applications and services?	2	0
		4	Is the data of statutory reporting requirements?	2	0
		5	Will the release of this data contribute to improve accountability?	2	0
		6	Could the data support decision making or emergency response activities at the individual, state, local, DOT or other external agency's level?	2	0
	Usability	1	Are the data entries clearly defined and described?	2.5	0
		2	Is the data consumer ready?	2.5	0
		3	Does the data have breadth of coverage (*for example,. national versus local*)?	2.5	0
		4	Is the frequency of data usage monitored?	2.5	0
	Availability	1	Does the data respond to need and demand as identified through public consultation?	2.5	0
		2	Is the dataset or tool available to the public with no access restrictions and at no cost?	2.5	0
		3	Is the data made available in appropriate amount of time?	2.5	0
		4	Is the data requested through FOIA on a recurring basis?	2.5	0
		5	Is the data currently publicly available on a DOT sponsored Web site, DOT publication, DOT report to congress or a DOT-sponsored research project?	2.5	0
		6	Is the data available in open format? (for example, CSV,XML,RSS,TXT,CMML,RTF,PDF, etc. For a complete list check the guidelines document.)	2.5	0
		7	Is the data currently available through a Web-based interface (ASP or other data mining tool)?	2.5	0
35 = Highest Quality	Completeness	1	Do you know who should be submitting the data and/or is there a data quality ordata steward point of contact?	3.5	0
		2	Is the dataset a full representation of the DOT data collection?	3.5	0

	Dimensions			Yes	No
35 = Highest Quality	Completeness	3	If the dataset is not a full representation, then what is the record submission percentage? (*Please choose the answer from the three options below.*) (number of records submitted or number of records that should have been submitted) * 100% (For example, measure missing rows from the dataset)		
			85%-99.99% = 3.5	3.5	0
			70% - 84.99% = 2.5	2.5	0
			<69.99% = 0	0	0
		4	If the dataset is not a full representation, then what is the fill rate on critical records? *(for example,. number of columns missing for every critical row in the dataset.)* *(Please choose the answer from the options below.)*		
			85%-99.99% =3.5	3.5	0
			70% - 84.99% = 2.5	2.5	0
			<69.99% = 0	0	0
		5	If a data is a subset of a larger collection, is there a risk of misinterpreting facts or the a possibility for invalid conclusions deriving from out-of-context data?	0	3.5
		6	Are computed data values documented and processes in place to ensure data integrity?	3.5	0
	Consistency	1	Does the database use DOT or other appropriate standards for data definitions?	3.5	0
		2	Are the data fields consistently represented? *(for example,. Contractor number with or without hyphen.)*	3.5	0
		3	Are data values consistently represented regardless of the context or time period in which they are presented?	3.5	0
		4	Do the data values comply with permissible values or business rules?	3.5	0
25 = Highest Manageability	Sensitivity	1	Would the release of data violate any current privacy requirements including OMB guidance and DOT guidance? *(For example,. FOIA Exemptions 6 & 7(C))*	0	5
		2	Would the release of data violate confidentiality following the NIST guidelines, OMB guidance and DOT guidance? *(For example,. FOIA Exemptions 4,7(D) & 8)*	0	5
		3	Does the data present a security risk at the data field level or in combination with other data? *(For example,. FOIA exemptions 1 & 7(F))*	0	5
		4	Is the data non public information that is for internal Government use in conducting its business or is prohibited from public disclosure? *(FOIA exemptions 2,5,7(A),7(E) & 7(B))*	0	5
	Level of Effort	5	Estimate the level of effort to convert the data to a releasable format: • People - *(Please choose the answer from the options below.)*		
			Less than 10 hours	2.5	0
			10-40 hours	2	0
			40-120 hours	1.5	0
			120-240 hours	1	0
			Greater than 240 hours	0	0

	Dimensions			Yes	No
25 = Highest Manageability	Level of Effort		• Resources: Equipment/Materials/Contracted Support - *(Please choose the answer from the options below.)*		
			Less than $3000	2.5	0
			$3001 - $25,000	2	0
			$25,001 - $100,000	1.5	0
			$100,001-$250,000	1	0
			Greater than $250,000	0	0
			Total Score		**/100**

Appendix B. Glossary

Term	Definition
Catalog	A collection of datasets
Data	Representations of facts, concepts, or instructions in a formalized manner suitable for communication, interpretation, or processing by human or automated means. The fundamental components of information
Data Element	A basic unit of identifiable and definable information that occupies the space provided by fields in a record or blocks on a form. A data element has an identifying name and value or values for expressing specific facts
Dataset	An organized collection of data, generally a compilation of data elements presented in tabular or other structured form with each column representing a particular variable and each row corresponding to a given value of that column's variable
Geodata	A specific kind of data pertaining to the geographic location and characteristics of natural or constructed features and boundaries on, above, or below the earth's surface, especially referring to data that is geographic and spatial in nature
Information	Any communication or representation of knowledge such as facts, data, or opinions in any medium or form, including textual, numerical, graphic, cartographic, narrative, or audiovisual form. Data processed in such a way that it can increase the knowledge of the person who receives it. Information is the output, or finished goods, of information systems
Metadata	Includes information that describes the characteristics of data, data or information about data, and descriptive information about an organization's data activities, systems, and holdings
Raw Data	for purposes of these guidelines, see entries **Data** or **Structured Data** or both
Structured Data	Data described via the E-R (Entity-Relationship) or class model, such as logical data models and XML documents. Structured data is organized in well-defined semantic sub-structures called entities
Tool	An interface through which a consumer may select a number of variables designed to filter a dataset. Results are presented to the user in the form the consumer desires (structured or unstructured). The consumer may then develop customized displays of the selected data
Unstructured Data	Data that is of a more free-form format, such as multimedia files, images, sound files, or unstructured text. Unstructured data does not necessarily follow any format or hierarchal sequence, nor does it follow any relational rules. Unstructured data generally refers to masses of computerized information which do not have a structure that is easily readable by a machine. Examples of unstructured data may include audio, video and unstructured text, such as the body of an email or word processor document