Montgomery County Government Technical Standards Manual for Publishing a Public Data Set

Department of Technology Services
Montgomery County, Maryland

VERSION	DATE	DESCRIPTION	AUTHOR
0.1	14 February, 2013	Initial Version	Mike Tarquinio
0.2	19 March, 2013	Edits	Mike Tarquinio, Dieter Klinger, Ivan Galic
1.0	13, May 2013	Comments from Socrata	Mike Tarquinio

Table of Contents

1.0 Introduction

1.1 dataMontgomery Program Description

The dataMontgomery program seeks to provide residents and constituents with direct access to County datasets in consumable formats. Providing this information offers the public an opportunity to review and analyze raw data, and the opportunity to use it for a variety of purposes.

1.2 Document Purpose

The purpose of this document is to describe a technical standard for the publishing of a public data set. This manual is required under Bill 23-12 [1] that was passed by the County Council and signed into law by the County Executive on 12/17/12.

The language in the bill that covers the creation and maintenance of a technical standards manual is:

2-157. Internet data set policy and technical standards.

(a) Within 180 days after this Article takes effect, the [[Department]] County must prepare and publish a technical standards manual for the publishing of a public data set in raw or unprocessed form through a single web portal by an agency to make public data available to the greatest number of users and for the greatest number of applications. The manual:

(1) must use open standards for web publishing and e-govemment, whenever practicable;

(2) must identify the reason why each technical standard was selected and to which types of data it applies;

(3) may recommend or require that data be published in more than one technical standard; and

(4) must include a plan to adopt or utilize a web application programmmg interface that permits application programs to request and receive public data sets directly from the web portal.

(b) The [[Department]] County must update the manual as necessary.

(c) The [[Department]] County must consult with appropriate voluntary consensus standards bodies and, when participation is feasible, in the public interest, and is compatible with agency and departmental missions, authorities, and priorities, participate with such bodies in the development of technical and open standards.

1.3 Document Change Management

The Montgomery County Government Technical Standards Manual for Publishing a Public Data Set document is published by the DTS Enterprise Architect. The Enterprise Architect is responsible for working with DTS Content Experts and department representatives to create a coherent Technical Standards Manual. The document adheres to stringent change management controls and follows a defined change management process.

Change requests can be initiated via DTS content experts, department representatives, the dataMontgomery workgroup, or the DTS Enterprise Architect. Contact the DTS Enterprise Architect, Mike Tarquinio (michael.tarquinio@montgomerycountymd.gov), for further details.

1.4 References

1. Montgomery County Government; *Bill 23-12, Montgomery County Open Government Data Act*; http://www6.montgomerycountymd.gov/content/council/pdf/bill/2012/20121204_23-12A.pdf*;* page accessed 1/15/2013
2. Montgomery County Department of Technology Services, July 19, 2007; *Enterprise Architecture Configuration Management Plan;*
3. Montgomery County Department of Technology Services, January 2, 2013; *Montgomery County Government Enterprise Architecture*; http://www.montgomerycountymd.gov/dts/architecture/index.html;
4. Montgomery County Department of Technology Services, January 2, 2013; *Montgomery County Government Enterprise Architecture Technical Architecture*; http://www.montgomerycountymd.gov/dts/resources/files/technicalarchitecture.pdf;
5. Montgomery County Government*; dataMontgomery*, *http://montgomerycountymd.gov/open/data.html;* page accessed 1/15/2013*;*
6. Montgomery County Government, December 12, 2012*; openMontgomery Montgomery County Maryland's Digital Government Strategy Building a 21st century program to better serve our residents, employees, and other partners;* http://montgomerycountymd.gov/open/Resources/Files/openMontgomery-Digital-Government-Strategy.pdf; page accessed 1/15/2013
7. Montgomery County Department of Technology Services, January 20, 2012*; Department of Technology Service Open Data Governance;* http://portal.mcgov.org/content/departments_intranet/dts/Services/OpenData/docs/open_data_governance.doc *;* page accessed 1/15/2013
8. Socrata, Inc; http://www.socrata.com; page accessed 1/15/2013
9. Socrata, Inc; *More time to build apps. Less time worring about the data*; http://www.socrata.com/solutions/socrata-for-developers/; page accessed 1/15/2013
10. Socrata, Inc*; Import Data Types; http://dev.socrata.com/publishers/import-data-types;* page accessed 1/15/2013

11. Executive Office Of the President, Office Of Management and Budget; *Open Data Policy-Managing Information as an Asset; http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf; page accessed 5/13/2013*

2.0 Data Principles

The guiding data principles for the dataMontgomery initiative are:

- **Complete** - All public data is made available. Public data is data that is not subject to valid privacy, security, or privilege limitations.
- **Timely** - Data is made available as quickly as necessary to preserve the value of the data.
- **Non-discriminatory** - Data is available to anyone, with no requirement of registration.
- **License-free –** Data is not subject to any copyright, patent, trademark, or trade secret regulation. Reasonable privacy, security, and privilege restrictions may be allowed.
- **Primary** – Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
- **Accessible** – Data is presented in a meaningful way.
- **Machine Processable** – Data can be processed.
- **Non-proprietary** – Data is available in a format over which no entity has exclusive control.

3.0 Data Standard

The dataMontgomery initiative is designed to publish County data in formats that the public can easily consume. The development of a data standard involves looking at the data as it exists in the County, as it rests on the delivery platform, and how it will be accessed by the public from the delivery platform.

The County has contracted with Socrata, Inc., a cloud-service provider, to provide the Montgomery County Open Data Platform at the website, https://data.montgomerycountymd.gov/. By using Socrata, the County is following Socrata standards for both storage and retrieval of information by the public. By using this product the County takes advantage of industry standard tools.

There are 3 general formats for the data as it travels to the public:

1. The first is the format of the data while it is internal to the County and before it is loaded onto the Socrata platform.
2. The second is the format of the data as it exists at rest on the Socrata platform.
3. The third is the format of the data external to Socrata when a user exports a data set from the County Open Data Platform. Upon export, the'general format' of the data can be a number of different downloadable formats, as well as a query string via the RESTful JSON API endpoint for that dataset.

3.1 Internal Format

When a data set is identified for publishing, the dataset owner must work with the Montgomery County Department of Technology Services (DTS) who is the County Socrata platform Administrator to define how the data will be prepared for upload to dataMontgomery. The first step is that a template for the data needs to be defined on the Socrata web site using Socrata supported data types. The next step is defining a strategy for taking the department's data set and transforming the data to match the Socrata template.

This design process creates a strategy for how the department's data will be transformed to meet the Socrata format. A design document will be created that formalizes the process so that the data can be reliably loaded each time. Additionally, by standardizing the initial upload the County will preserve the ability to use a different vendor in the future.

As part of the design process, the department and DTS must carefully review their data in order to represent it in the most valuable form for an end user. All data elements must be described with the description including the range of possible values for the data element. Best practices for the data transform process are:

- data should be broken down into component parts rather than being a large unstructured string
- data should be geocoded
    - preference is for lat/long information but address information can also be used
- unstructured types like plain or formatted text should be avoided and more structured data types like numbers, dates and times, phone, etc should be used

- data should be in a raw, unsummarized format
- data should, whenever possible, be in csv to avoid including excel formula columns and associated workbooks
- footers and citation info should be removed from the data prior to upload

By following the above guidelines and using more detailed data types, the transform process that Socrata uses when presenting data to the user will provide a more useful transform type. For example, Socrata can provide more informative XML tags that describe the data when the user selects an XML format for the output.

3.1.1 Socrata Template

Before data can be uploaded to Socrata a template must be created. Socrata supports a row/column format where the columns are named, described, and a type associated with it. Socrata supports the following types in its import process:

- Numbers, Money, and Percent
- Dates and Time
- Checkboxes
- Emails
- Website Links/URLs
- Location Columns
- Plain Text
- Formatted Text
- Photo (Image)

After the import process Socrata supports the following that can be added as a new column:

- Multiple Choice
- Document
- Nested Table
- Dataset Link

The next section contains details for some of the supported data types. Section 3.1.2 was copied verbatim from the Socrata Import Data web site located at: *http://dev.socrata.com/publishers/import-data-types*. The description was accurate on the date it was retrieved which was 2/10/2013.

3.1.2 Socrata Import Data Format

========== **begin Socrata-import-data-types** ==========

Numbers, Money, and Percent

For numbers we directly use Java's BigDecimal parsing. For details see the Java documentation.

A percent can be either a number preceded or followed by a percent (%) sign or just a number. Percentages aren't in the range 0.0 to 1.0 like they are in Excel. A percentage input of 42.0 is idiomatically 42.0%.

Money can be either a number preceded with a dollar sign ($ -- more currency symbols soon) or just a number. For negative monetary values, either a negative sign or a set of parentheses are acceptable: e.g. $-42.21, ($42.21), -$42.21 or (42.21).

Dates and Times

Dates are parsed by default in the American/Pacific (PST) timezone. You can explicitly specify a timezone by using the supported ISO 8601 subset. A Z character is UTC, otherwise the offset is [+-]HH:mm.

For inputs that don't specify a time of date, the resulting time is undefined. In other words, don't rely on it being anything consistent.

The accepted input formats (ISO and non-ISO) are:

Supported ISO 8601 Subset

- yyyy-MM-dd[T]HH:mm:ssZ (e.g. "1920-01-22T00:00:00Z", "1920-01-22T00:00:00-10:00", or "1920-01-22 00:00:00Z")
- yyyy-MM-dd[T]HH:mm:ss (e.g. "1920-01-22T00:00:00" or "1920-01-22 00:00:00")
- yyyy-MM-dd[T]HH:mm (e.g. "1920-01-22T00:00")
- yyyy-MM-dd (e.g. "1920-01-22")

Supported non-ISO Dates

For dates other than the ISO subset we accept a date, optionally followed by a time, i.e.

(date)[(time)]

Non-ISO dates are always parsed in the American date format locale (i.e. month, day, year). Months and days can be either single or double digit and may or may not be led with a '0'. Years can be either four digits (preferred) or two. If a year is two digits it will be assumed to be between 1951 and 2050: i.e. 1/2/75 would be January 2nd 1975, but 1/2/49 would be January 2nd 2049.

The accepted input formats are:

- MMM d, yyyy (e.g. "Jan 4, 1982")
- MMM d, yy (e.g. "Jan 4, 82")
- MMMM d, yyyy (e.g. "January 4, 1982")
- MMMM d, yy (e.g. "January 4, 82")
- M-d-yyyy (e.g. "1-4-1982")

- M/d/yyyy (e.g. "1/4/1982")
- M.d.yyyy (e.g. "1.4.1982")
- M-d-yy (e.g. "1-4-82")
- M/d/yy (e.g. "1/4/82")
- M.d.yy (e.g. "1.4.82")

Checkboxes

Valid false values:

- 0
- f
- false
- n
- no
- off

Valid true values:

- 1
- t
- true
- y
- yes
- on

Emails

Three different input formats are acceptable for emails.

1. Sam Gibson
2. sam.gibson@socrata.com
3. Sam Gibson <sam.gibson@socrata.com>

Nearly all emails should work, though technically for performance' sake we only support a subset of the RFC regex for emails. If there's a specific email or set of emails that's causing you a problem, please feel free to submit a support ticket and we'll fix it.

Website Links/URLs

URL's support two different input formats. Only three URL schemes are acceptable: ftp, http, and https. We use a custom regular expression to validate URLs. It should accept just about anything that you throw at it, but there's always a chance that it's missed something.

1. <a href"http://www.socrata.com/">Socrata
2. http://www.socrata.com/

Location Columns

Since Location columns are a "composite" column that's created by appending multiple values together, they can only be created if the data in the matching column is formatted in the correct manner.

1. To format a latitude and longitude pair to be appended or refreshed, format the values in the given column as: (xx.xxx, yyy.yyy) where xx.xxx is the latitude and yyy.yyy is the longitude. Make sure that your values are in decimal degrees, and that you use "negative" longitude degrees for "degrees west" (ie. "-122.36" for Seattle, WA, and "2.33" for Paris, France).
2. To append or refresh an address, simply format it as a comma separated US-formatted address within the column, such as 101 Yesler Way, Seattle, WA, 98108. It'll automatically be queued up for geocoding if the address parser recognizes the format.
3. You can also specify a latitude and longitude pair along with your address: 101 Yesler Way, Seattle, WA, 98108 (47.60165, -122.33403).

Whatever file format you use, make sure that the values in your location column are properly escaped. For example, for CSV, an address must be wrapped in double-quotes in order to escape the commas within it:

"(47.60165, -122.33403)"

========== **endn Socrata-import-data-types** ==========

3.1.3 Upload Process

There are two methods for upload with one being manual and the other being automated. In either case, DTS will work with the Using Department to do the actual upload.

A key point with either process is the creation of a design document. The design document will document the transform that must take place mapping the internal County data to the Socrata format. The purpose of the design document is to define a repeatable process that will take the data to the Socrata platform.

The automated process involves using the DTS Enterprise Service Bus (ESB). The ESB has the ability to take input from a number of data sources with some examples being spreadsheets, email, web services, files and databases. The ESB has a rich transform capability and can take most structured data and transform it into an acceptable input for Socrata. More detailed information on the ESB capabilities can be found in the document Montgomery County Government Enterprise Architecture Technical Architecture [4] in the Service Enabled Domain.

3.2 Data at Rest

The Data at Rest format is going to be the vendor's format. Currently, the vendor is Socrata and the Socrata format is the standard. They have a documented format as shown in section 3.1.1.

3.3 External Format

Access by the public to a dataset is standardized through the Socrata vendor offering. Information around the supported access can be found on the Socrata developer website at http://dev.socrata.com/. Socrata provides an API to retrieve the data called the Socrata Open Data API (SODA). SODA is described as "an open, standards-based RESTful application programming interface" [9]. With the Socrata platform every dataset is automatically provided with a simple REST API that a developer can use to download the dataset.

By using the REST API for the particular dataset, the developer can access the dataset in any of the supported data formats:

- JSON
- XML
- CSV
- RDF
- XSL
- XSLX
- PDF
- RSS

The Socrata platform will perform a transform from their at rest data format to the requested data format.

More detailed information on the SODA API can be found at: http://dev.socrata.com/. The Socrata developer website contains the following information and resources:

- dynamic code samples
- interactive developer console with which the developers can explore the API
- step by step guides
- native libraries for Ruby, Java, Python, JavaScript, C# and PHP
- community forums
- detailed documentation
- videos
- developer blog with Socrata support

Developer resources can also be found at http://github.com/socrata.

The Socrata API supports a number of industry standard protocols. It is licensed under the Creative Commons Attribution 3.0 license. Terms of the license can be found at http://creativecommons.org/licenses/by/3.0/.