2014

Open Data Standard and Technical Standards Manual

State Chief Information Officer Sean Vinck

State of Illinois

3/26/2014

Table of Contents

1.0 Executive Summary

Public Act 98-0627 (PA 98-0627 or the Act) was enacted on March 10, 2014 to further increase transparency and accountability in government by establishing a new State Open Operating Standard. The Act encourages state agencies and local governments across Illinois to increase the amount of raw data they share with each other, businesses, academic institutions and the general public. The State of Illinois Open Data site, www.data.illinois.gov, is a searchable clearinghouse of information from State Agencies that is helping inform citizens about the operation of State government and encouraging the creative use of State information, including the development of applications for mobile devices that can be built around the data. By May 2014, the Office of the Governor, through the State Chief Information Officer (State CIO) is to prepare and publish a technical standards manual that will establish policies and standards for open data.

PA 98-0627 sets forth the following timeline:

- On or before May 9, 2014, the Office of the State Chief Information Officer, on behalf of the Office of the Governor, must prepare and publish a **technical standards manual** for the publishing of public data sets in raw or unprocessed form through a single web portal by State Agencies for the purpose of making public data available to the greatest number of users and for the greatest number of applications and shall whenever practicable, use open standards for web publishing and e-government.
- On or before May 9, 2014, the Office of the State Chief Information Officer, on behalf of the Office of the Governor, must prepare and publish **portfolio management policies** for ensuring compliance with the requirements of the Act.
- On or before July 8, 2014, each State Agency must submit a **compliance plan** and a **draft longer term strategic enterprise application plan** consistent with PA 98-0627 to the Office of the Governor and shall make such **plan available** to the **public on the web portal**.

This document is the Technical Standards Manual (TSM) that defines the statewide policies, standards, and guidelines required by the Act.

2.0 Introduction

2.1 Background

In June 2011, the State of Illinois Open Data Portal ("**Open Data Portal**"), www.data.illinois.gov, was created as an integral part of the Illinois Innovation Council's plan to promote economic growth through innovation and the engagement of citizens, developers, academia and industry. Currently, the site hosts data from multiple State Agencies as well as federated data from the Federal website www.data.gov, and contains more than 6,400 data sets. The Open Data Portal presents a powerful tool for innovators to create value-added assets.

On March 10, 2014, the Illinois General Assembly unanimously passed PA 98-0627, a bill introduced by the Governor's Office and intended to increase transparency, accountability and savings in government by establishing a new State Open Operating Standard.

The purpose of the Technical Standards Manual ("**TSM**") is to identify the reasons why each technical standard was selected and the types of data for which it is applicable. The TSM includes a plan to adopt or use web application programming interface that permits application programs to request and receive public data sets directly from the single web portal. The TSM includes the common and repeated use of rules, conditions, guidelines, or characteristics for products or related processes and production methods, and related management systems practices and: (i) the definition of terms; (ii) classification of components; (iii) delineation of procedures; (iv) specifications of dimensions, materials, performance, designs, or operations; (v) measurement of quality and quantity in describing materials, processes, products, systems, services, or practices; (vi) test methods and sampling procedures; or (vii) descriptions of fit and measurements of size or strength.

2.2 Definitions

Several terms in the Technical Standards Manual have specific meanings as defined in the Act and these are included in the list below. Terms in this TSM take their plain language meaning unless otherwise stated in this Definitions Section. **The meaning of any given term in this TSM is not necessarily the same as the meaning of that same term in the Act**.

1. "**Agency**" means any office, administration, department, division, bureau, board, commission, advisory committee or other government entity performing a governmental function of the State of Illinois.
2. "**Compatibility**" means the ability of two or more systems or components to seamlessly perform required services; the capability of a computer, device, or program to function with or substitute for another make and model of computer, device, or program; or the capability of one computer to run the software written to run on another computer.

3. "**Data**" means statistical or factual information, accompanied by units of measurement whenever possible, in alphanumeric form existing as a list, chart, graph, table or other non-narrative form recorded as a measurement, transaction or determination related to the mission, or business operations, of an Agency.
4. "**Data format**" means the organization of information for display, storage or printing. Data is maintained in certain common formats so that it can be used by various programs, which may only work with data in a particular format, e.g. .pdf, HTML.
5. "**Data Set**" or "**Dataset**" means a named collection of related records on a storage device, with the collection containing individual data units organized or formatted in a specific and prescribed way, often in tabular form, and accessed by a specific access method that is based on the data set organization.
6. "**Ecosystem**" means the interrelationship of various open data participants — the State, the Agencies, Open Data Portal, software applications, providers of public data, policies, technical standards, and services that make the Open Data Standard work and be useful for end-users.
7. "**Machine-readable**" (also called **computer-readable**) refers to information encoded in a form which can be scanned or sensed by a machine or computer and interpreted by its hardware and software.
8. "**Metadata**" (**Meta data**) means structured data that contain or define other data, making it more discoverable in online environments. Metadata also provides context to data and can make data easier to reuse and combine with other data. Metadata can also include information about the quality of the data.
9. "**Metadata tag**" means a tag accompanying each piece of data describing the attributes, provenance, and required security protections of that piece of information.
10. "**Open format**" means one that is platform independent, machine readable, and made available to the public without restriction that would impede the re-use of that information.
11. "**Open Operating Standard**" means a set of detailed technical guidelines used as a means of establishing uniformity in an area of hardware or software which is collaboratively developed, clearly defined and recognized by an independent body. Open Operating Standard is the preferred method of performing a single function or a number of interrelated functions in a uniform manner that is vendor neutral and encourages interoperability by not being confined to any one platform.
12. "**Public Data**" means all data that is collected by any unit of State or local government in pursuance of that entity's official responsibilities which is otherwise subject to disclosure pursuant to the State's Freedom of Information Act (FOIA), 5 ILCS 140/ et. seq., and is not prohibited from disclosure pursuant to any other contravening legal instrument, including but not limited to, a superseding provision of Federal or state law or an injunction from a court of competent jurisdiction.
13. "**Public Data Set**" means a comprehensive collection of interrelated data that is available for inspection by the public in accordance with any provision of law and is maintained on a computer system by, or on behalf of, an Agency, excluding any data to which an Agency may deny access pursuant to any provision of law or any federal or state rule or regulation.

14. "**Raw unstructured data**" means data that requires manual review or manipulation and are not structured for automatic processing by a computer system. Data is often qualitative rather than quantitative.
15. "**Unstructured data**" means computerized information, which does not have a data structure. This may include audio, video and unstructured text such as e-mails or documents; or defined data that does not reside in fixed fields of a database; e.g., word processing documents, email, and other non-database records.
16. "**State**" means State of Illinois.
17. "**Statewide Standard**" means an industry standard or de facto standard which is adopted and/or mandated by the Office of the Governor through the State CIO to be used in relation to the State's information technology systems and services. The Office of the Governor, via the State CIO, will establish statewide standards that apply to all State Agencies.
18. "**Structured data**" means sources, which represent a collection of records stored in a computer in a systematic way, with each record organized in a definitive schema, describing the objects that are represented in the database and the relationships among them.
19. "**Voluntary consensus standards body**" means a domestic or international organization which plan, develop, establish, or coordinate voluntary consensus standards using agreed-upon procedures. A voluntary consensus standards body is defined by the following attributes: (i) openness, (ii) balance of interest, (iii) due process, (vi) an appeals process, and (v) consensus, which is defined as general agreement, but not necessarily unanimity, and includes a process for attempting to resolve objections by interested parties, as long as all comments have been fairly considered, each objector is advised of the disposition of his or her objection(s) and the reasons why, and the consensus body members are given an opportunity to change their votes after reviewing the comments.

2.3 State of Illinois Open Data Portal

The Open Data Portal is a clearinghouse of various data sets in a standard format that is readable by virtually all computer systems. Development of the Open Data Portal is aimed at increasing access to public data or the data used in the operations of government, and making it accessible to the citizens of Illinois in a conventional method. The Open Data Portal can be accessed via www.data.illinois.gov and currently utilizes the Socrata Open Data Application Programming Interface and platform.

2.4 Data Assessment and Publication Frequency

Each State Agency shall assess its data as often as is reasonable. Each Agency, following internal data assessment, shall upload new data and, if applicable, revised data to the Open Data Portal as often as is reasonable but in no case less often than as required by the policies, procedures and protocols established by the Office of the Governor.

3.0 Technical Standards

This Section instructs each Agency subject to PA 98-0627 and any unit of local government that elects to publish its data to the Open Data Portal how to:

(i) determine which data formats are compatible with the Open Data Portal;
(ii) determine how to translate its unstructured data into an Open Data Portal-compatible format;
(iii) generate and prioritize metadata;
(iv) evaluate which data is not subject to the requirements of the Act and/or is prohibited from disclosure; and
(v) upload data subject to the requirements of the Act to the Open Data Portal.

3.1 Data Formats Compatible with the Open Data Portal

Agencies should import their raw unstructured data into one or more data sets. The raw unstructured data should not contain any coding languages such as HTML tags.

The raw unstructured data may contain the following data types:

1. ***Numbers, money, and percentages***

For numbers, use the Java's BigDecimal parsing. For details, see http://download.oracle.com/javase/1.5.0/docs/api/java/math/BigDecimal.html#BigDecimal(java.lang.String)

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).

2. ***Dates/Times***

ISO 8601 is the International Standard for the representation of dates and times. ISO 8601 describes a large number of date/time formats. For more details, see http://www.w3.org/TR/NOTE-datetime

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, 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")

3. ***Booleans***

Valid false values:

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

Valid true values:

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

4. ***Email addresses***

The input format for email addresses should be firstname.lastname@domain.gov

5. ***URLs***

Only three URL schemes are acceptable: ftp, http, and https. We use a custom regular expression to validate URLs. Do not include any HTML formatting tags such as <script>, <table>, <tr>, <td>, or
.

http://www.illinois.gov/

6. ***Location columns***

Since Location columns are a "composite" column that's created by appending multiple values together, they can't be created at import, but they can be appended or refreshed into 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 values are in decimal degrees, and to 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 will automatically be queued up for geocoding if the address parser recognizes the format.
3. Also specify a latitude and longitude pair along with the address: "101 Yesler Way, Seattle, WA, 98108 (47.60165, -122.33403)".

Whatever data format used, make sure that the values in the data 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.

3.2 Translation of Unstructured Data to Open Data Portal

The translation of unstructured data into one or more data formats for compatibility for the Open Data Portal may be imported using the following formats:

Element	Type	Description
CSV	Text	Comma-separated values
TSV	Text	Tab-separated values
XLS	Binary	Microsoft Excel
XLSX	Binary	Microsoft Excel XML

Agencies can use other methods to import unstructured data into one or more data formats for compatibility for the Open Data Portal.

3.3 Generation and Prioritization of Metadata

Accurate, specific metadata shall be generated for each data set for which metadata is able to be generated. The process of generating metadata is beneficial to the Agencies as a means of internal organization and self-evaluation. Descriptive, accurate metadata benefits third parties to consume data on the Open Data Portal.

Metadata shall be presented in plain language while capturing all elements of the uploaded data set to which it applies. When composing metadata, Agency employees should assume that third party consumers of that metadata's corresponding data do not have an understanding of that Agency's operations; therefore, all metadata shall be composed as basically and clearly as possible while erring on the side of providing an abundance of explanation rather than an insufficient amount.

All elements and variables of each data set shall be clearly explained. Units of measurement must be provided for all data elements to which they apply. No agency may upload a data set knowing additional data not contained in that data set is within the Agency's possession and control; in any case where an employee of an Agency is aware of additional data not in that Agency's possession or control that would fully complete or make more complete that Agency's uploaded data set, that fact must be conspicuously stated in the corresponding metadata.

Pursuant to PA 98-0627, for purposes of prioritizing data sets, agencies must consider whether information contained therein:

(i) can be used to increase agency accountability and responsiveness;
(ii) improves public knowledge of the agency and its operations;
(iii) furthers the mission of the agency;
(iv) creates economic opportunity;
(v) is received via the online forum for inclusion of particular public data sets; or
(vi) responds to a need or demand identified by public consultation.

Data containing any of these elements shall be prioritized for upload to the Open Data Portal. If an Agency's resources required to upload data to the Open Data Portal are limited such that not all data can be uploaded at the frequency required by section 2.4 of this document, publication shall be made in a hierarchical manner according to the order of the six priorities listed in the immediately preceding paragraph.

3.4 Data Not Subject to PA 98-0627

Refer to the definition of "data" in PA 98-0627 for a list of data not subject to the requirements of that Act nor to the requirements of this technical standards manual.

Data not subject to the requirements of the Act include:

1. data to which an agency may deny access pursuant to any provision of a federal, state or local law, rule or regulation;

2. data that contains a significant amount of data to which an agency may deny access pursuant to any provision of a federal, state or local law, rule or regulation where redacting such protected data in order to publish the unprotected elements would impose undue financial or administrative burden;

3. data that reflects the internal deliberative process of an agency or agencies, including but not limited to negotiating positions, future procurements, or pending or reasonably anticipated legal or administrative proceedings;

4. data stored on an agency-owned personal computing device, or data stored on a portion of a network that has been exclusively assigned to a single agency employee or a single agency owned or controlled computing device;

5. materials subject to copyright, patent, trademark, confidentiality agreements or trade secret protection;

6. proprietary applications, computer code, software, operating systems or similar materials;

7. employment records, internal employee-related directories or lists, facilities data, information technology, internal service-desk and other data related to internal agency administration; and

8. any other data the publication of which is prohibited by law.

3.5 Publication of Data to the Open Data Portal

Please refer to the Open Data Portal's import specifications for details on formatting and parsing of data types, and on publishing data. These specifications and instructions can be found by following this link:

http://dev.socrata.com/publishers/import-data-types

4.0 Public Technical Standards

The purpose of this section is to explain the support for third party applications and application programming interfaces for data to be published in multiple types of formats.

4.1 Web Application

The Illinois Open Data Portal supports most modern web browsers. The web application permits the listing, viewing, exporting, embedding, filtering, visualizing, personalizing, commenting on, and rating of public data sets.

4.2 Application Programming Interface (API)

The Illinois Open Data Portal supports a generic API that permits access to all published data sets in a similar manner. The API is based upon the Socrata Open Data API (SODA).

4.3 Download Formats

In 2009, the U.S. Government defined an open file format as "one that is platform independent, machine readable, and made available to the public without restrictions that would impede the re-use of that information."

The Illinois Open Data Portal supports the public downloading of data sets in the following formats:

Format	Type	Description
CSV	Text	Comma-separated values
JSON	Text	JavaScript Object Notation
PDF	Binary	Portable Document Format
RDF	Text	Resource Description Framework
RSS	Text	RDF Site Summary/Really Simple Syndication *(Note: each row is represented as a separate item, but each item's description field contains an HTML table with the column names and row values)*

XLS	Binary	Microsoft Excel
ZIP	Binary	Typically contains a shapefile set (SHP, SHX, DBF) (Geographic data sets only), or a collection of multiple files which are all part of the same data set. *(Note: this is usually provided within a compressed archive)*