OPEN NEW YORK

NEW YORK STATE

OPEN DATA HANDBOOK

data.ny.gov

November 6, 2013

Office of Information Technology Services

Contents

Figures

"Our state government possesses vast treasure troves of valuable information and reports: from health, business and public safety data to information on parks, recreation, labor, and transportation … The Open New York web portal will allow researchers, citizens, business and the media direct access to high-value data, which will be continually added to and expanded, so these groups can use the data to innovate for the benefit of all New Yorkers."

Governor Andrew M. Cuomo, State of the State Address, January 9, 2013

Introduction

Open Data – data.ny.gov

Quality by Design

- New York State is committed to proactively releasing publishable state data.
- New York State is committed to making publishable state data openly and freely available in accessible formats for the public to re-use and consume.
- New York State is committed to publishing high quality data with comprehensive metadata and documentation to foster interoperability and maximize citizen understanding of the data.
- New York State is committed to on-going and continuous publication of publishable state data.

Government is the public's business and the public should have access to the records of government. New technologies have dramatically changed both the way government conducts business and the public's expectations about access to government information. As part of this transformation, New York State launched Open.ny.gov, on March 11, 2013, dedicated to increasing public access to one of the State's most valuable assets – data. Its goals are to spark innovation, promote research and economic opportunities, engage public participation in government, increase transparency, and inform decision-making.

Concurrent with the launch of Open NY, Governor Andrew Cuomo issued Executive Order No. 95, "Using Technology to Promote Transparency, Improve Government Performance and Enhance Citizen Engagement." This unprecedented Executive Order directed covered state entities, for the first time, to identify and catalogue their data, and make publishable state data available on the new transparency website.

The concept of “Open Data” describes data that is freely available, machine readable, and formatted according to uniform technical standards to facilitate visibility and re-use of published data. New York's open data platform is a quality by design web-based public data portal that catalogues data and enables data to be discoverable. The portal offers access to standardized data that can be easily retrieved, downloaded, sorted, searched, analyzed, and re-used by citizens, business, researchers, journalists, developers, and government to process, trend, and innovate utilizing a singular dataset or combinations of datasets.

OpenNY puts tools for transparency, accountability, and innovation directly into the hands of New Yorkers and people all around the world through a centralized, user-friendly interface. This increased visibility provides derivative value as the public is able to analyze and utilize government data, and better understand what is happening in government on all levels – federal, state, and local.

The Open Data Handbook

This Open Data Handbook is intended as a general guide for government entities participating in data.ny.gov[1], as well as the general public. The Handbook provides guidelines for identifying, reviewing, and prioritizing publishable state data for publication – with a foundational emphasis on high quality, and metadata and documentation requirements. These guidelines are intended for use by both covered state entities and other government entities not covered by Executive Order 95 (including localities). These guidelines are also intended for use by the public in order to understand how New York State makes its publishable data sets available.

The breadth of data and agency participation are continually being enhanced and expanded on data.ny.gov, making it a dynamic, living platform. This Handbook, providing guidelines for publication of publishable State data onto data.ny.gov, is the first step in a major shift in the way New York State government agencies share information to promote efficiency, accessibility and transparency; a major shift in the way New York State government engages citizens and fosters innovation and discovery in the scientific and business communities. It begins the process of standardizing the state’s data, which will make it easier for both government workers and the public to discover and use the data. This, in turn, advances “interoperability” so the data can be more easily shared and analyzed.

Working in collaboration with state agencies, this Handbook will be supplemented, as needed, with technical and working documents addressing specific formatting, data preparation, and

[1] For ease of use, the Handbook will refer to participating government "agencies," although that term encompasses other government organizational structures that may participate in Open NY, such as government authorities.

data refresh and data submission requirements. We welcome input from academics, researchers, developers, businesses, entrepreneurs, and the general public to help identify data that would be useful to them and the best ways of providing it.

Timeline

In advancing a transparency agenda, New York State is committed to continuous publication of publishable State data. Data discovery is an iterative and on-going process. The following is a high-level timeline for the implementation of Executive Order 95.

March 11, 2013	Executive Order No. 95 issued
March 11, 2013	Open.ny.gov website launched
April 10, 2013	Agency Data Coordinators Appointed
April 25, 2013	Data Working Group established
June 10, 2013	ITS issues Provisional Open Data Handbook
September 9, 2013	Covered State entities create catalogues of publishable data; covered State entities' propose schedule to ITS for making publishable state data publicly available.
November 6, 2013	ITS issues Final Open Data Handbook
April 1, 2014	ITS begins publishing quarterly reports
December 15, 2019	Open Data and compliance with Executive Order No. 95 is fully incorporated into state covered entities' on-going core business planning and strategies.

Definitions

Throughout this handbook, some terms are used as defined in Executive Order 95. These definitions are included below.

Covered State Entity

(i) Any State agency or department, or any office, division, bureau, or board of such State agency or department, except where the head of such agency or department is not appointed by the Governor,

(ii) any State board, committee, or commission, at least one of whose members is appointed by the Governor, and

(iii) all public-benefit corporations, public authorities and commissions, for which the Governor appoints the Chair, the Chief Executive, or the majority of Board Members, except for the Port Authority of New York and New Jersey.

Chief Data Officer (CDO) The New York State Chief Data Officer in the Office of Information Technology Services or a designee thereof.

Data Final versions of statistical or factual information that:

(i) are in alphanumeric form reflected in a list, table, graph, chart or other non-narrative form, that can be digitally transmitted or processed;
(ii) are regularly created or maintained by or on behalf of a covered State entity and are controlled by such entity; and
(iii) record a measurement, transaction or determination related to the mission of the covered State entity.

The term "data" shall not include image files, such as designs, drawings, photos or scanned copies of original documents; provided, however, that the term "data" shall include statistical or factual information about image files and geographic information system data.

Data set A named collection of related records maintained on a storage device, with the collection containing data organized or formatted in a specific or prescribed way, often in tabular form.

ITS The New York State Office of Information Technology Services.

Publishable State data Data that is collected by a covered State entity where the entity is permitted, required or able to make the data available to the public, consistent with any and all applicable laws, rules, regulations, ordinances, resolutions, policies or other restrictions, requirements or rights associated with the State data, including but not limited to contractual or other legal orders, restrictions or requirements. Data shall not be Publishable State data if making such data available on the Open Data Website would violate statute or regulation (e.g., disclosure that would constitute an unwarranted invasion of personal privacy), endanger the public health, safety or welfare, hinder the operation of government, including criminal and civil investigations, or impose an undue financial, operational or administrative burden on the covered State entity or State.

Directives for Participating Agencies

The following directives are based upon Executive Order 95.

Open Data Website

New York State Office of Information Technology Services will create an Open Data Website for the purposes of collecting and dissemination Publishable State data.

Data Coordinator

Each covered State entity will designate a Data Coordinator. The Data Coordinator shall:

- have authority equivalent to that of a Deputy Commissioner or the head of a division or department within their agency;
- have knowledge of data and resources in use by their agency; and
- be responsible for their agency's compliance with the Executive Order, this handbook, and future directives which may be needed to support the open data program.

The Data Coordinator serves as the liaison between the ITS data.ny.gov team and the agency. As such, the Data Coordinator is best positioned to convey to their agency any specific needs of the Open Data platform (such as formatting the data or defining a structure that is optimal for publication). Insofar as agencies vary in terms of size, functions, and complexity, larger agencies may also identify individuals within specific divisions, bureaus, and/or units (as data champions) to assist the agency data coordinator.

Data Working Group ("DWG")

ITS will form a Data Working Group ("DWG"). The DWG will assist the CDO in carrying out his or her duties under this Order. The DWG is made up of representatives from ITS and the Information Security division of ITS, the New York State Office of General Services, the Division of Budget, a local government expert from the Department of State and eight to twelve Data Coordinators, who represent an appropriate cross-section of covered State entities.

Open Data Handbook

ITS, in consultation with the DWG, is required to issue guidance in this Open Data Handbook to covered State entities about implementing the Executive Order. Within 90 days after the date of the Executive Order, ITS shall issue a provisional Open Data Handbook. Within 240 days, ITS shall issue a final Open Data Handbook. The Open Data Handbook may be amended by ITS from time to time.

Publication of Data

Each covered State entity shall create a catalogue of their Publishable State data, and propose a schedule to ITS and the CDO for making its Publishable State data publicly available. Such schedules shall be made publicly available and provide for updating the data catalogue as appropriate. Each covered State entity shall prioritize data publication in accordance with guidelines set forth in the Open Data Handbook.

Participation by Localities and Other State Entities

Localities are invited, and are encouraged, to submit data to the Open Data Website for publication in accordance with guidelines set forth in this Open Data Handbook.

New York State agencies and authorities other than covered State entities shall be permitted, and are encouraged, to submit data to the Open Data Website for publication in accordance with guidelines set forth in the Open Data Handbook.

Publishing Approvals

Participating agencies are required to engage in an internal review process and obtain approvals for the datasets which the agency wishes to commit to the Open Data Website. Agencies are responsible for driving towards increasing data content quality and accuracy, and are responsible for ensuring compliance with all security, privacy, confidentiality laws, rules, and regulations, as well as any Intellectual Property Rights requirements and status under the NYS Freedom of Information Law (including whether data may lawfully be withheld under FOIL's limited exceptions).

For any particular dataset, at a minimum, agencies must receive explicit approval and sign-off from the individuals listed below. Standardized approval forms provided by ITS must be completed and signed prior to dataset publication. Agencies may determine additional internal approvals and signatures are required, and should include such additional persons in their review and sign off process (e.g. Public Information Officer)

The Data Coordinator is responsible for obtaining the following approvals from within their agency:

- **Data Owner:** This is typically the head of an agency department, a bureau director, or person situated similarly within the agency and likely to have been directly involved with the collection of the data. The Data Owner will have the greatest familiarity with and knowledge of the dataset and the data it contains, and the purpose for the collection of the data. The Data Owner should know the accuracy and currency of the data, and be best able to describe and fill in the metadata elements describing the data. Approval by the Data Owner also validates that the agency has obtained permission and knowledge from the department which is most responsible for the specific data. (This may also be referred to as Program owner, or Data Steward)

- **Legal counsel (e.g. in-house or an outside attorney where applicable**): Legal counsel will likely be in the best position to determine whether the dataset has internally been reviewed sufficiently to ensure compliance with privacy and security requirements, intellectual property rights, and FOIL responsibilities. It is recommended that the legal counsel consult with the agency's chief privacy official, chief security officer, FOIL officer, and records access officer.

- **Point of Contact: (i.e., the agency's Data Coordinator):** The Point of Contact is the liaison between data.ny.gov and the agency.   This person is best positioned to convey to the Data Owner any specific needs of the data.ny.gov platform maintainers for the data to be formatted or defined in an optimal manner for publication. (For example, an individual dataset may need to be cleansed to remove extraneous, non-machine-readable elements). This person also serves as an additional internal control ensuring the dataset has been properly evaluated before being provided to data.ny.gov.

- **Chief Executive or designee**: Approval by the head of the agency ensures full knowledge within the agency and that the agency is providing a dataset to data.ny.gov under full authority to do so. It also serves as the ultimate internal control to exercise authority within the agency to ensure proper evaluations of the datasets have been completed.

Standardization

The way data consumers interact and use the Open Data Platform is greatly influenced by the way the data is published. The Open Data Platform requires the agencies to present the data in a machine-readable format to enable software tools, applications & systems to process it. Instead of preprocessing the data, data consumers can directly access the raw data and customize the data for their own consumption needs.

Standardizing the data publishing model on the platform in a machine-readable format enables automation leading to the development of new, friendly analysis tools. The same data can be reused for another business use case without extra processing.

As part of the standardization process, the Open New York Data platform has identified the following minimum requirements:

Metadata

The platform will support a common and fully described core metadata scheme for each hosted dataset and Application Program Interface (API) within the data catalogue. The metadata scheme would allow data publishers to classify selected contextual fields or elements within their dataset as well as adhere to common Meta attributes identified platform-wide empowering the data consumers to build automated discovery mechanisms at a granular-level. Using a common metadata taxonomy will allow Open New York to convey and increase discoverability of high-value datasets

data.ny.gov adheres to core components of the Dublin Core standard for metadata (http://www.dublincore.org/documents/dces/).   The ability to search and find information is enhanced by the adherence to metadata standards required with each dataset. In addition, metadata is linked to subject categories which provides for more precise searching and document management. Adoption of the Dublin Core, together with standards for Open Data, maximizes adaptability and interoperability.

The Dublin Core Metadata Initiative (DCMI) is incorporated as a non-profit organization hosted at the National Library Board of Singapore. Its lists of elements, glossary, and FAQs were last revised in 2005, but an effort to update its User Guide is being developed at the wiki page http://wiki.dublincore.org/index.php/User_Guide. data.ny.gov uses the current set of elements, which are required to accompany each dataset (Refer to Appendix B: Metadata Elements for additional details).

Descriptive Information

data.ny.gov serves as a platform to present machine-readable data, so that end-users may process, access, discover, extract and combine data elements to discover new insights, observations, and utility regarding the data. In furtherance of New York State's commitment to high quality, data.ny.gov requires agencies to submit metadata and supplemental documentation with each dataset (e.g., data dictionaries, overview documents, etc.). This ensures data are fully described to maximize the public's understanding and interpretation of the data and facilitates interoperability.

Domain Categories

The platform supports a common domain model that allows data publishers to identify, transform and anchor datasets in a particular domain. Using a common scheme based on categories and tags, Open New York standardizes the available domains within the platform, thus helping data consumers to retrieve datasets readily and uniformly using either the standard core metadata or extending the search using domain-specific metadata. Refer to Appendix B for examples of categories.

Catalog Sharing

data.ny.gov currently federates with a number of data catalogues. The ability to federate in a bi-directional way creates a nimble and agile data ecosystem facilitating interoperability. As technology advances, and publication of data worldwide continues to grow exponentially, there is a wealth of metadata being assembled. The ability to query embedded metadata across catalogues can generate invaluable insights and, as such, data.ny.gov will explore common, open formats (such as DCAT).

Data Sets

In order to facilitate automatic processing of the data, make it easily accessible and available in machine-readable formats, standardized open data file formats for data publishers and data consumers must be made available so that they can be uploaded, retrieved, indexed and searched.

Open Specifications

Published data sets must be compatible with open specifications, where they exist, and support the requirements of the agency and its data and adaptability to facilitate interoperability. For example, additional open formats may be included, over time, such as (KML/KMZ) and GeoJSON.

Content Formats

Datasets published to the Open New York platform are machine-readable and have a clear separation of metadata from the original source data. Data Consumers will be able to automate the process of discovery, accessibility and ingestion by using the uniform open data formats supported by the platform.

Tabular Data

The current Open Data Platform supports the following formats:

- **CSV & TSV**: Comma/Tab Separated Values

Geographic Data

Geospatial data is usually organized as a collection of features that define a layer. Layers can be overlaid on top of one another, allowing visualization spatial relations, spatial queries, and analysis. The Open Data Platform supports two data formats for geospatial information. The appropriate format is dependent on the specific characteristics of the underlying geographic data.

- **Points**: All Tabular File Formats or Shapefile
- **Lines**: Shapefile
- **Polygons**: Shapefile

Point data can be stored in either tabular or shapefile format. Tabular formatting of points requires either columns for latitude and longitude, or complete address information (house number, street, village/town/city, state, and zip code) that can be geocoded. In contrast, lines and polygons define complex geometric structures that are not easily defined as column attributes. Therefore, shapefile format is a preferred format for these complex geographic structures.

Each shapefile (at a minimum) should contain the following files:

- **.shp**: Defines the geometry (shapes)
- **.dpf**: Defines the attribute table
- **.prj**: Projection, ensures the feature locations are accurately rendered on the map
- **.shx**: Shape indexing file, for efficient processing

Note: Shapefiles which use projections other than WGS-1984/Web Mercatur will require conversion which may result in a minimal loss of accuracy. In some cases this conversion can be handled by the Open Data Platform; in other cases it must be done by the participating agency.

Other supported geospatial formats may include Keyhole Markup Language (KML/KMZ).

Geocoding

The Open Data New York Platform supports geocoding services which convert human-readable address information into mappable coordinates (Latitude/Longitude).

Updates to Published Data Sets

Data sets on data.ny.gov must be kept up-to-date. Specific guidance regarding updates will be addressed in technical and working documents previously referenced. Four mechanisms are supported for refreshing a dataset.

- Replace: All existing records are removed and new records are inserted.
- Append: New dataset records are inserted
- Update: Existing records are modified
- Delete: Existing records are removed

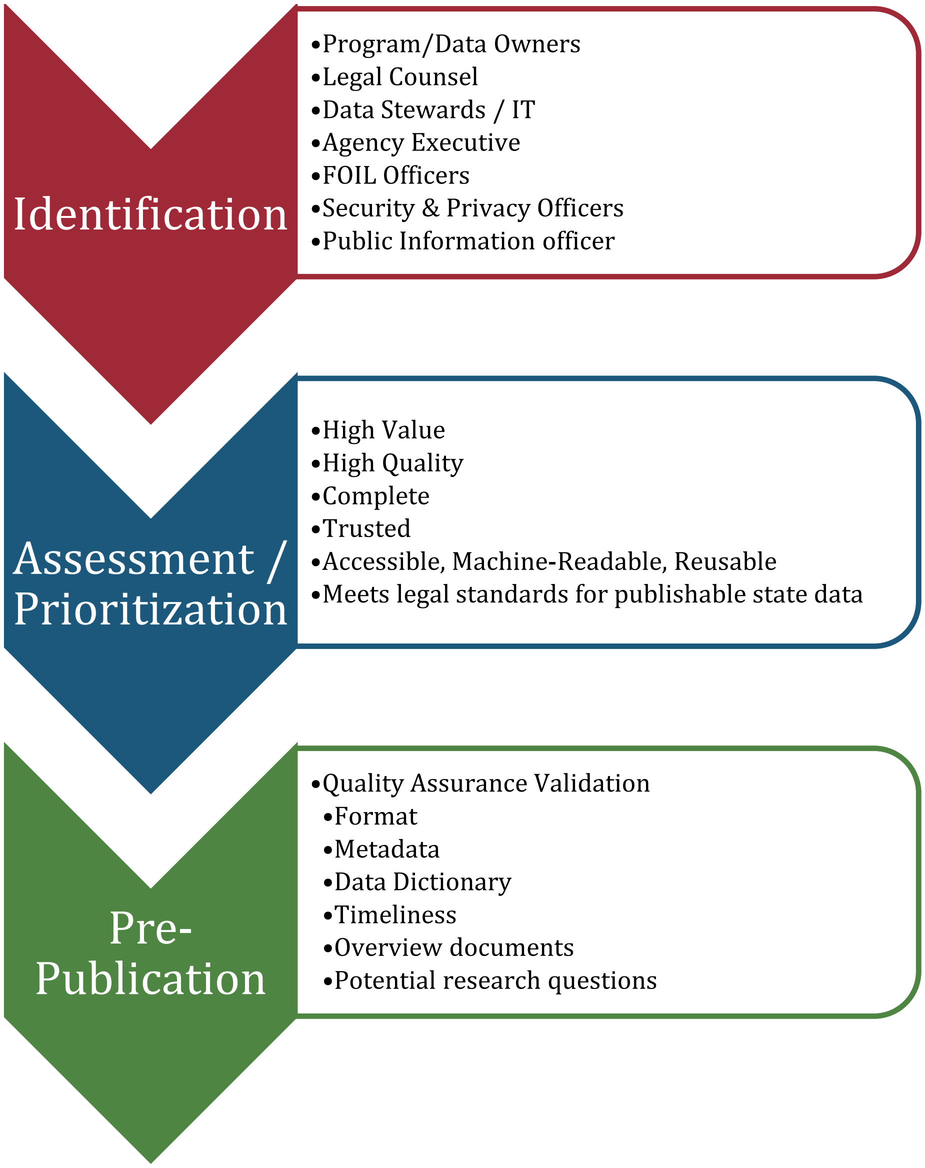
The frequency of updates is included in the metadata each provides with each dataset (see metadata posting frequency).

Guidelines for Participating Agencies

Executive Order No. 95 provides a specific definition of "Publishable State Data" to guide covered State agencies. Publishing data on data.ny.gov involves a collaborative multi-step agency process (see Figure 1: Guidance Summary). In identifying Publishable State Data, agencies should include analyses from their executive and program staff, data coordinators, FOIL officers, data stewards/IT, public information officers, security and privacy officers, and legal counsel.

Covered State entities (and entities not covered by Executive Order 95) vary widely in terms of size, personnel, functions, responsibilities, mission, and data collected and maintained. As such, the identification and prioritization processes may vary across agencies and entities. These guidelines serve to provide assistance across a broad spectrum of agencies, with the stipulation that agencies look to their governing laws, rules, regulations, and policies in identifying and publishing "publishable state data."

Figure 1: Guidance Summary



Identification

- Program/Data Owners
- Legal Counsel
- Data Stewards / IT
- Agency Executive
- FOIL Officers
- Security & Privacy Officers
- Public Information officer

Assessment / Prioritization

- High Value
- High Quality
- Complete
- Trusted
- Accessible, Machine-Readable, Reusable
- Meets legal standards for publishable state data

Pre-Publication

- Quality Assurance Validation
  - Format
  - Metadata
  - Data Dictionary
  - Timeliness
  - Overview documents
  - Potential research questions

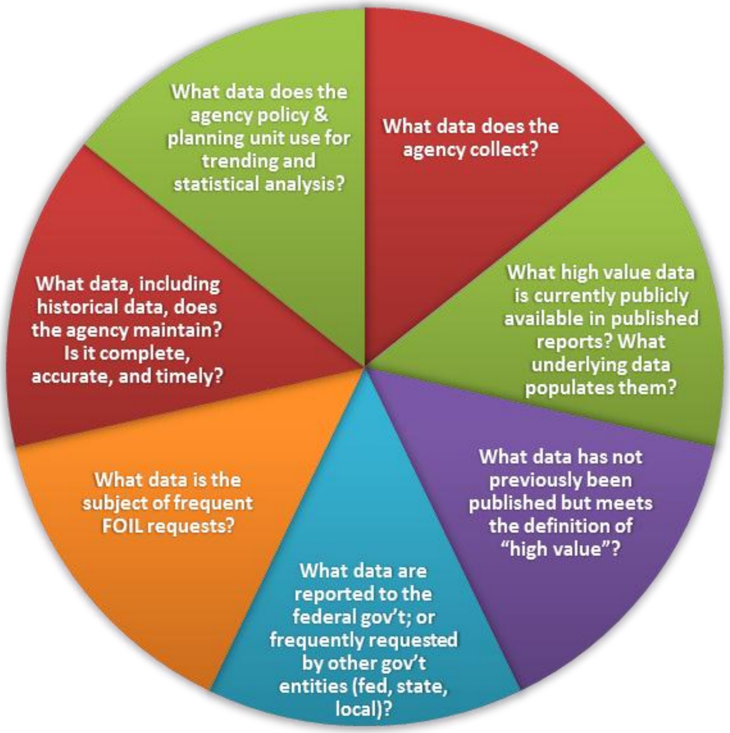
Data Set Identification

In creating a data catalogue, agencies should identify those datasets that are high value, high quality, complete, and in accordance with the definition of "Publishable State Data" within Executive Order 95. “High value” data, as defined within Executive Order 95, is that which can be used to increase the agency’s accountability and responsiveness, improve public knowledge of the agency and its operations, further the mission of the agency, create economic opportunity, or respond to a need or demand identified after public consultation.

The questions in Figure 2, and below, are neither exhaustive nor may be applicable to all agencies, but serve to provide a framework to identify potential data for publication on data.ny.gov.

For each question, agencies must assess whether the data falls within the definition of "Publishable State Data" and the disclosure considerations that follow.

Figure 2: Identifying Publishable Data Sets


What data does the agency collect?
What high value data is currently publicly available in published reports? What underlying data populates them?
What data has not previously been published but meets the definition of "high value"?
What data are reported to the federal gov't; or frequently requested by other gov't entities (fed, state, local)?
What data is the subject of frequent FOIL requests?
What data, including historical data, does the agency maintain? Is it complete, accurate, and timely?
What data does the agency policy & planning unit use for trending and statistical analysis?

General Questions

- **What data (including historical data) does the agency collect, maintain, or hold?** Conducting a preliminary survey of data can be an excellent first step towards identifying publishable state data.
- **What "high value" data are currently publicly available?** Agencies already publish a considerable amount of data online, but it may not necessarily be accessible in bulk, or available through machine-readable mechanisms. Reviewing weekly, monthly, or quarterly reports which are frequently accessed by the public, or public-facing applications, which allow visitors to search for records, are excellent starting points.

- **What underlying data populates aggregate information in published reports?** Published reports are often populated with data which is compiled or aggregated from internal systems. For example, a weekly public report may indicate that an agency has closed 25 projects in that week. The internal system, which has details of each case, may have additional details which can be made public.
- **What data does the agency policy and planning unit use for trending and statistical analysis?** Similar to published reports, trend and statistical analysis is often performed using data from various sources. Those sources can be reviewed for data which can be made public.
- **What data are reported to the federal government; or frequently requested by other government entities (federal, state, local)?** Reporting of data, which is required by statute, grant, or other agreement, may already be done by an agency. Reviewing these reports (and their underlying data sources) can help identify data which can also be provided to the public. In addition, meeting these reporting requirements (particularly statutory ones) might be accomplished simply by making the data set(s) available on data.ny.gov.
- **What data is the subject of frequent FOIL requests? What data is the public requesting?** There are multiple methods by which the public requests data from agencies. For example, some Freedom of Information Law (FOIL) requests may seek to obtain datasets or records which are to be provided back to the requestor in digital format. These requests (particularly repeated requests for the same dataset) might be fulfilled by making the dataset(s) available on data.ny.gov.
- **What data have not been previously published but meet the definition of "high value"?** Publishable state data that can be used to increase the covered State entity's accountability and responsiveness, improve public knowledge of the entity and its operations, further the mission of the entity, create economic opportunity, or respond to a need or demand identified after public consultation.

Do the datasets represent discrete, usable information

In identifying datasets, government entities may be concerned that users of data.ny.gov will not understand their raw data or, if distilled to its rawest form, might lose utility. For example, state and local rules might differ, such that publishing raw, separate datasets of the two may reduce the value of the raw data being combined into a single dataset.

There are no hard and fast rules about what level of detail is sufficiently granular to add value to a government dataset. Whenever possible, government entities should resist the temptation to limit datasets to only those the agency believes might be understood or useful. Entities should be wary of underestimating the users of data.ny.gov. data.ny.gov users may come from a variety of fields and specialties, including academic and other government users who can envision a use for the raw data not anticipated by the originating entity. A better practice is for the agency to ensure its metadata describing the dataset is complete, including comprehensive overview documents describing the data, data collection, data fields, and presentation of research questions to maximize the utility and usefulness of the data.

Release Prioritization

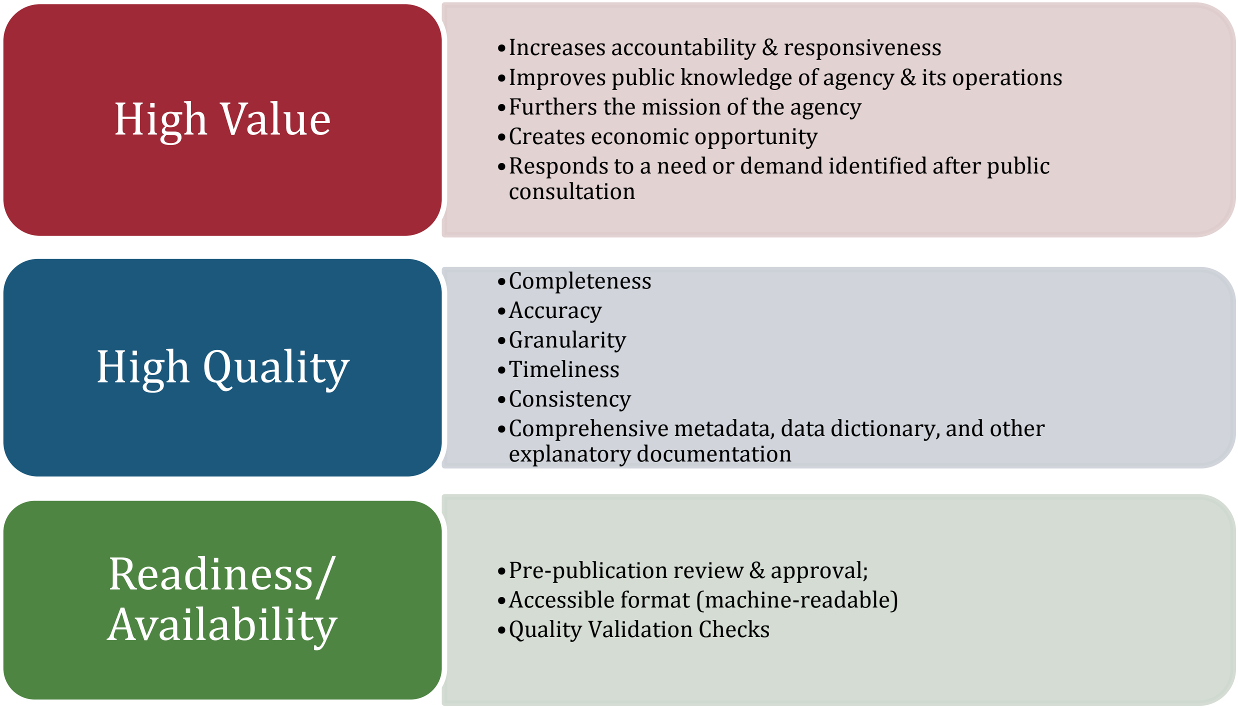
Executive Order 95 states: *"Prioritization of publication of data based on the extent to which the data can be used to increase the covered State entity's accountability and responsiveness, improve public knowledge of the entity and its operations, further the mission of the entity, create economic opportunity, or respond to a need or demand identified after public consultation..."*

Executive Order 95 further states: *"Data shall not be Publishable State Data if making such data available on the Open Data Website [data.ny.gov] would…impose an undue financial, operational or administrative burden on the covered State entity or State."*

When creating a schedule for publication of a particular dataset, agencies must make an assessment based upon a number of different factors. Agencies may use the guidance below to determine the priority for each data set. Prioritizing initial and ongoing publication will entail balancing high value with data quality, data availability, and data readiness. Each covered State entity shall create schedules and prioritize data publication in accordance with guidelines set forth herein, and in a timely manner, recognizing that it may take time for agencies to prepare high quality data (noting that datasets vary in complexity and, as such, can significantly vary in preparation time).

In prioritizing data for release, therefore, agencies must account for time to: identify data, assess the data (i.e., ensure consistency, timeliness, relevance, completeness, and accuracy of the data), ensure completeness of the metadata and data dictionary, review and obtain all necessary approvals to publish the data, and prepare data, metadata and requisite accompanying documentation for publication (Figure 3).

Figure 3: Prioritization



High Value

- Increases accountability & responsiveness
- Improves public knowledge of agency & its operations
- Furthers the mission of the agency
- Creates economic opportunity
- Responds to a need or demand identified after public consultation

High Quality

- Completeness
- Accuracy
- Granularity
- Timeliness
- Consistency
- Comprehensive metadata, data dictionary, and other explanatory documentation

Readiness/ Availability

- Pre-publication review & approval;
- Accessible format (machine-readable)
- Quality Validation Checks

Below are suggested questions, the answers to which can assist agencies in prioritizing publication of high value "publishable state data" consistent with Executive Order 95:

i. **Does the data highlight agency performance, or might publication of the data benefit the public by setting higher standards?** The agency might be in the forefront of standards for government performance, where exposing the data might cause other agencies to raise their performance**.**

ii. **Has the data ever been published or made available in a machine-readable format so that it can be processed, analyzed, or re-used?** There may exist procedures in place which can be leveraged to publish the data, such as exports for periodic department reviews, or routine exchanges of data with other agencies.

iii. **Is the data "high value?"** While "high value" can be subjective, your agency best understands the needs of the constituency that it serves. Publishing relevant data can ultimately support those needs.

iv. **Does availability of the data align with new State and/or Agency initiatives?** The ordering publication of any relevant datasets accordingly might be of great value.

v. **Does availability of the data align with federal initiatives or exposures of federal data?** There may be higher value in the agency's data if synergies can be created.

vi. **Can publication of the data address regulatory or grant requirements?** While some data required by regulations or grants may be inappropriate for public release, publishing the data may be an acceptable way to meet those requirements and make the data accessible to the public simultaneously.

vii. **Does the data support decision making at the state, local, internal agency or other external agency's level, or contain information that informs public policy?** Publishing such a dataset publicly can be a powerful platform for fostering productive civic engagement and policy debate.

viii. **Is the data timely? What is the dataset refresh and maintenance cycle?** Systems, which support the ongoing operations of an agency, are often kept up-to-date on a daily basis. Publishing raw or aggregate data drawn from these systems can provide tremendous value.

ix. **Does availability of the data align with legal requirements for data publication?** For example, there might be statutorily-required reporting which can be satisfied by publishing datasets, without necessarily needing an extensive narrative report. If the data is collected and compiled by the agency to fulfill statutory reporting requirements, then the agency's governing laws have already determined that the data is of high value for that agency.

x. **Would availability of the data improve agency-to-agency communication?** Certain government functions may involve multiple agencies requiring access to similar data.

xi. **Could availability of the data create specific economic opportunity?** In many cases, this will be unknown to the agency in advance. Some of the greatest successes of the open data movement have involved government data being commercially appropriated in useful ways, such as weather data. To the extent the agency can anticipate significant commercial use of the data, the agency may wish to order publication of such data more highly as it creates its schedule.

xii. **Could the data be useful for the creation of novel and useful third-party applications, mobile applications, and services?** Software applications often leverage data from multiple sources to provide value to their customers. Making agency data sets available can support the delivery of greater value (and impact) through those applications.

xiii. **Does the data further the core mission or strategic direction of the agency or multiple government entities?** Publishing aggregated data (statistics, metrics, performance indicators) as well as raw data can often help an agency advance its strategic mission. In addition, data.ny.gov can serve as a conduit for efficiently sharing information with other agencies.

xiv. **Does the data have depth and breadth of years of coverage?** Release of data with high information content and quality can improve accountability and responsiveness and/or improve public knowledge of the agency and its operations.

xv. **Does the data have accompanying metadata and a data dictionary**? Metadata and all accompanying documents should be comprehensive so as to provide a full understanding of the data and data elements to an end-user. This ensures version control, availability of contact information, and descriptive information sufficient for end-users to be able to use and interpret the data. In addition, where applicable, agencies should append disclaimers to highlight limitation of the data and/or prevent use of the data in misleading ways.

xvi. **Is the data accurate/complete?** The dataset must be sufficiently final or complete, such that it is currently publishable. Agencies should work to transform any data sets or partial data sets which are not complete or high quality so that they can eventually be published. If there is a trigger allowing the agency to publish the data at some time in the future, then scheduling publication of the data should be set accordingly.

xvii. **Is the dataset in a format that is machine-readable or can be easily transformed?** The data should be organized or formatted in a manner which is machine-readable and that can be re-used, and capable of being digitally transmitted or processed. It should be in tabular or geo-spatial form. Agencies should consider the level of effort required to transform the data to a machine-readable format and maintain it in such a format.

xviii. **Is the data frequently requested?** As demand is known and quantifiable, this should raise the value of this data for publication. If the dataset is the type that is requested through FOIL on a recurring basis, then the agency may reduce duplication and obtain efficiencies by posting data on data.ny.gov.

xix. **Is the data needed by the public after-hours?** As demand may be known and quantifiable. Generally when there is this type of demand for the data, such datasets should be ranked, where applicable, of higher value.

xx. **Does the data have a direct impact on the public?** The data is likely of higher value if it is already apparent there is a deep impact and interest by the public (e.g., hospital infection rates, food establishment inspection results, etc.).

xxi. **Is the data in strong demand from constituencies?** The data might be of higher value to specific, narrow interest groups which may be the agency's core constituency for those issues.

xxii. **Is the data of timely interest?** Announcements of progress or success – or reactions to public criticism - can be strongly supported by publishing related data, should it exist.

Disclosure Considerations

As agencies classify data sets and catalogue Publishable State Data, they should be mindful of legal and policy restrictions on publication of certain kinds of data. The following guidelines regarding disclosure provide additional factors for consideration as agencies begin to identify and review datasets.

i. **Security, Privacy, Regulatory, & Aggregate Data.**
The public release of some agency data might result in the violation of laws, rules, or regulations. Some data may not be appropriate to release because it can compromise internal agency processes, such as procurement. Other data may contain personally

identifiable information. Finally, even if detailed data appears innocuous, it may be possible to easily combine it with other public information to reveal sensitive details (commonly known as the *mosaic* effect).

Even if there are no *legal* impediments to publishing the data, releasing the data may have unintended or undesirable effects. For example, posting anonymized arrest records on a weekly basis might inadvertently reveal where police are concentrating enforcement efforts.

ii. Thresholds

Various statutes and regulations, such as Health Insurance Portability and Accountability Act (“HIPAA“) and its privacy regulations, have very exacting requirements for determining whether data have been sufficiently de-identified so as not to compromise individual privacy. For example, the presence of medical conditions per geographic location might constitute high-value, useful, and sought-after data; however, exposing it might identify individuals and their medical conditions.

Another example is the Family Educational Rights and Privacy Act of 1974 (FERPA). Under FERPA, the Federal Government has established guidelines for data privacy to prevent individuals from being identified indirectly from aggregation of data. Agencies that deal with student educational data should be aware of guidelines that restrict publication of some data.

Even in the absence of specific legal prohibitions, government entities should watch for outlier publication conditions. For example, identifying a single arrestee who is a minor of a certain age in a certain county without providing any other information, might nonetheless serve to identify that particular individual.

For particular datasets that pose such issues agencies may consider providing aggregated data based upon their laws, rules, regulations, and policies. Alternatively, agencies may set disclosure thresholds for the dataset (many agencies already adhere to such standards). For example, if a cell in a particular dataset field goes below a certain number of individuals, the value in that particular cell should be hidden. Government entities will need to balance their desires to publish accurate, complete, and valuable tabulations against the need to guard against unwarranted invasions of personal privacy, in specific situations.

iii FOIL Applicability

Under the NYS Public Officers Law, Article 6 (the NYS Freedom of Information Law, or "FOIL"), the presumption is that government records shall be open to the public, unless excludable under a narrow set of specific exemptions including such concerns as invasion of personal privacy, impairment of contractual or collective bargaining negotiations, exposure of protected trade secrets, interference with law enforcement or judicial proceedings, endangering life or safety, and others. Government entities should confer with their FOIL officers for publication of data on data.ny.gov, and exclude any datasets which, because their publication would cause the harms described in the FOIL law, would not constitute "Publishable State Data."

iv Ownership Rights

In some circumstances, an agency may not possess all the necessary rights to be able to publish a specific data set. For example, if the data was collected or compiled by a third party, there may be a contractual or intellectual property limitation which prevents it from being made public. In these cases, the appropriate permission must be secured from the sourcing entity, and additional disclaimers may be required.

Narrative Content

Narrative content on its own is not appropriate for publishing on data.ny.gov. However, such content may have been developed based upon existing agency data which has already, or will be, published.

If an agency develops extensive narrative reports about published data, then those reports should be published on the agency's website, while providing a link to the published data set on data.ny.gov. It is important to keep this link current.

Public Use of data.ny.gov

The following section is intended for use by the public when interacting with data.ny.gov.

Public Input

Data.ny.gov facilitates citizen engagement and participation directly through the website. The portal provides the capability by which the public can engage with the data provider, and accepts submissions from users through the "Give Feedback" form. In addition, data.ny.gov provides a mechanism for the public to submit and request online, through the site’s “Suggest a Dataset,” datasets not currently released.

Downloading Bulk Data

The Open New York Data platform provides an open, standards-based, RESTful application programming interface[2] to provide automatic access to the publicly published datasets within the open data catalogue. The platform will be optimized towards the developer community, technical users, and researchers and enable programmatic reuse.

Datasets

The Open Data Platform supports two classifications of datasets: tabular and geospatial. A tabular dataset is a flat file that conforms to a predefined schema. The schema defines the characteristics of a fixed number of columns, including the column name and data type. A geospatial dataset contains information that can be readily rendered on an underlying map. Examples of geospatial features include points (buildings), polylines (bus routes), and polygons (school districts), along with attribute information that describes characteristics of each spatial feature.

Tabular

Datasets can be exported for download in popular human-readable formats, machine-readable standards and streamable file formats. The Open Data Platform currently supports the following exportable tabular file formats:

- **CSV**
- **JSON**
- **PDF**
- **RDF**
- **RSS**
- **XLS**
- **XLSX**
- **XML**

Large Files

Public data often consists of historical archives, comprised of potentially millions of records collected over an extended period of time. The Open Data Platform supports the loading, exporting and visualization of large datasets (> 1GB).

Geo-Spatial

Geospatial data contains geographic features and attribute data that defines the properties of geographic features. Attributes are stored in a tabular format with unique key references to

[2] An Application Programming Interface, or "API," is a set of routines, protocols, and tools for building software applications that enable software components to interact with each other. See API section below. "REST" (Representational State Transfer) is a type of software architecture for networked, distributed systems.

the associated geographic features.   Two export methodologies are supported for exporting geographic information: geospatial and attribute. Attribute layers can be exported as tabular data export file formats as identified above under Tabular.

Geospatial data can be downloaded in any of the tabular formats defined above, as well as the following formats:

- **Shapefile**
- **Keyhole Markup Language (KML/KMZ)**

Application Program Interface (API)

All data on the platform can be consumed via an Application Program Interface (API) which provides direct access to the raw data for developers.   The platform's APIs allow the end user to get results back in JSON, XML, RSS, etc. This separation of data model and encoding allows the support of many different encoding standards, even ones that do not yet exist.

This enables users to access the data in a host of different file formats that is independent of the original format of the data.

Access to Datasets

The Open New York Platform supports the existence of an API Strategy that allows the developer community to dynamically query a dataset within the data catalogue. Each hosted dataset within the data catalogue will:

i) be readily and uniformly accessible.
ii) be available for automated processing by applications and systems
iii) have a standard API Endpoint

The Endpoint points to a RESTful implementation of the underlying dataset that has been accessed. All communication with the API is done through an HTTPS protocol. The platform provides the following preferred response types which are made available by specifying the format as part of the URI, or by sending the appropriate HTTP "Accepts" header:

- **JSON**
- **XML**
- **CSV**
- **RDF**

Featured API Catalog

Additionally, the Open New York Platform supports the creation of a featured API Catalog that provides custom endpoints to the developer community to dynamically query the raw dataset based on "specified" dataset elements. Just like published data sets, the featured API Catalog is categorized and tagged using the common domain and metadata schema

Terms of Use

The OPEN-NY Terms of Use may be found on the OPEN-NY website at the following link: https://data.ny.gov/download/77gx-ii52/application/pdf. The terms are subject to modification as conditions warrant. When the Terms of Use change, this will be indicated within the Terms them-selves with notification of the "Last Modified Date.

Appendix A: Executive Order 95

EXECUTIVE ORDER 95

USING TECHNOLOGY TO PROMOTE TRANSPARENCY, IMPROVE GOVERNMENT PERFORMANCE AND ENHANCE CITIZEN ENGAGEMENT

WHEREAS, the State possesses vast amounts of valuable information and reports on all aspects of life in New York State, including health, business, public safety, and labor data as well as information on transportation, parks, and recreation; and

WHEREAS, new information technology has dramatically changed the way people search for and expect to find information, and such technology can aggregate ever larger quantities of data and allow government to provide information to the public with increasing efficiency; and

WHEREAS, the State can use these powerful tools to enhance public access to government data and make government in New York State more transparent in order to promote public trust; and

WHEREAS, ensuring the quality and consistency of such data is essential to maintaining its value and utility;

NOW, THEREFORE, I, Andrew M. Cuomo, Governor of the State of New York, by virtue of the authority vested in me by the Constitution and laws of the State of New York, do hereby order as follows:

A. Online Website. An online Open Data Website for the collection and public dissemination of Publishable State data, and, to the extent feasible, reports is hereby established. The Open Data Website shall be maintained at data.ny.gov or such other successor website maintained by, or on behalf of, the State, as deemed appropriate by the New York State Office of Information Technology Services in consultation with the Governor's Office and Data Working Group established below. The Open Data Website will provide "single-stop" access to Publishable State data that is owned, controlled, collected or otherwise maintained by covered State entities as defined herein and, to the extent feasible, reports of such covered State entities.

B. Definitions. As used herein, the following terms shall have the following meanings:

1. "Covered State entity" means (i) any State agency or department, or any office, division, bureau, or board of such State agency or department, except where the head of such agency or department is not appointed by the Governor, (ii) any State board, committee, or commission, at least one of whose members is appointed by the Governor, and (iii) all public-benefit corporations, public authorities and commissions, for which the Governor appoints the Chair, the Chief

Executive, or the majority of Board Members, except for the Port Authority of New York and New Jersey.

2. "Chief Data Officer" shall mean the New York State Chief Data Officer in the Office of Information Technology Services or a designee thereof;

3. "Data" shall mean final versions of statistical or factual information that (i) are in alphanumeric form reflected in a list, table, graph, chart or other non-narrative form, that can be digitally transmitted or processed; (ii) are regularly created or maintained by or on behalf of a covered State entity and are controlled by such entity; and (iii) record a measurement, transaction or determination related to the mission of the covered State entity. The term "data" shall not include image files, such as designs, drawings, photos or scanned copies of original documents; provided, however, that the term "data" shall include statistical or factual information about image files and geographic information system data.

4. "Data set" means a named collection of related records maintained on a storage device, with the collection containing data organized or formatted in a specific or prescribed way, often in tabular form.

5. "ITS" shall mean the New York State Office of Information Technology Services.

6. "Publishable State data" shall mean data that is collected by a covered State entity where the entity is permitted, required or able to make the data available to the public, consistent with any and all applicable laws, rules, regulations, ordinances, resolutions, policies or other restrictions, requirements or rights associated with the State data, including but not limited to contractual or other legal orders, restrictions or requirements. Data shall not be Publishable State data if making such data available on the Open Data Website would violate statute or regulation (e.g., disclosure that would constitute an unwarranted invasion of personal privacy), endanger the public health, safety or welfare, hinder the operation of government, including criminal and civil investigations, or impose an undue financial, operational or administrative burden on the covered State entity or State;

C. Open Data Website Administration

1. The Open Data Website shall be administered by ITS.
2. The Chief Data Officer ("CDO") and the Chief Technology Officer within ITS shall coordinate implementation and expansion of the Open Data Website to facilitate the sharing of information and initiatives resulting from developments based on this Order.
3. Within 30 days after the date of this Order, each covered State entity shall designate a Data Coordinator, who shall: (i) have authority equivalent to that of a Deputy Commissioner or the head of a division or department within the covered State entity; (ii) have knowledge of data and resources in use by the entity; and (iii) shall be responsible for that covered State entity's compliance with this Order.

4. Within 45 days after the date of this Order, ITS and the CDO shall establish a Data Working Group ("DWG") made up of representatives from ITS and the Information Security division of ITS, the New York State Office of General Services, the Division of Budget, a representative from the Department of State with expertise in local government and at least eight but no more than twelve Data Coordinators, who shall represent an appropriate cross-section of covered State entities. The DWG shall assist the CDO in carrying out his or her duties under this Order.

D. Publication of Data. All covered State entities shall make their Publishable State data available on the Open Data Website as follows and in accordance with the Open Data Handbook to be promulgated by ITS:

1. Each covered State entity shall create a catalogue of their Publishable State data within 180 days after the date of this Order.
2. Each covered State entity shall, within 180 days after the date of this Order, propose a schedule to ITS and the CDO for making its Publishable State data publicly available. Such schedules shall be made publicly available and provide for updating the data catalogue as appropriate.
3. Each covered State entity shall create schedules and prioritize data publication in accordance with guidelines set forth in the Open Data Handbook.

E. Opportunity for Localities to Participate. Localities are invited, and are encouraged, to submit data to the Open Data Website for publication in accordance with guidelines set forth in the Open Data Handbook. ITS shall assist localities so they may use the Open Data Website. Such assistance shall include, but not be limited to, technical assistance and expertise, and accommodations shall be made for variations among local governments' capacity and equipment.

F. Open Data Handbook. ITS, in consultation with the DWG, shall issue guidance to covered State entities on implementing this Order in the Open Data Handbook.

1. The Open Data Handbook shall:
    1. provide models and guidelines for covered State entities to follow when creating their data catalogues;
    2. provide guidance to covered State entities on setting a schedule for initial and ongoing publication of data on the Open Data Website including but not limited to requiring:
        1. consultation with the directors and staff of the covered State entity's public affairs or public information, legal and Freedom of Information Law ("FOIL") offices;

        2. prioritization of publication of data based on the extent to which the data can be used to increase the covered State entity's accountability and responsiveness, improve public knowledge of the entity and its operations, further the mission of the entity, create economic opportunity, or respond to a need or demand identified after public consultation;

    3. provide guidelines for identifying and reviewing publishable State data by covered State entities before publication;

    4. provide uniform standards for the format of data submitted for publication on the Open Data Website;

    5. set forth the Open Data Website terms of use;

    6. provide guidelines on participation by agencies and authorities other than covered State entities and participation by localities;

    7. provide guidance on the publication of narrative data, such as reports; and

    8. set forth any further definitions and guidance necessary for the implementation of this Order.

2. Within 90 days after the date of this Order, ITS shall issue a provisional Open Data Handbook. Within 240 days after the date of this Order, ITS shall issue a final Open Data Handbook.

3. The provisional and final Open Data Handbook shall be made public by ITS and the CDO, including through the Open Data Website. ITS shall solicit and consider comments and suggestions related to the Handbook from State agencies, authorities, localities and the public.

4. The Open Data Handbook may be amended by ITS from time to time.

G. Notification to the Public of Ongoing Publication of Data. The public shall be notified of additions and updates to the data catalogue contained on the Open Data Website.

H. Participation by Other State Entities. New York State agencies and authorities other than covered State entities shall be permitted, and are encouraged, to submit data to the Open Data Website for publication in accordance with guidelines set forth in the Open Data Handbook.

I. Covered State entities and all other participating agencies, authorities and localities shall not be prevented from publishing data in advance of the dates set in their schedules if the data has been approved for publication by ITS.

GIVEN under my hand and the Privy Seal of the State in the City of Albany this eleventh day of March in the year two thousand thirteen.

BY THE GOVERNOR

Secretary to the Governor

Appendix B: Metadata Elements

Metadata Element	Description	Dublin Core Value
Dataset Name / Title	The name of the dataset as it will appear on the platform.	Title
Dataset Description	Short description that explains the purpose of the Dataset and the data within	Description
Category	The general category that the dataset is included in on the site (The categories include: Economic Development, Education, Energy & Environment, Government & Finance, Health, Human Services, Public Safety, Recreation, Transparency, and Transportation).	Type
Tags / Keywords	Keywords about the dataset used for searching purposes.	Subject
Data Provided By	The Agency that provided the data.	Contributor
URL to Dataset Program Web Page	The URL to the program area web pages.	n/a
Responsible Organization Within Agency	The organization that the dataset owner is a part of	Creator
Time Period	The timeframe of data available in the associated data file (e.g., Beginning 2005).	Coverage (temporal)
Create Date	The date the resource was made available in its present form – auto generated	Date
Posting Frequency	How often the Dataset will be refreshed (e.g., Annually, Monthly, Daily).	n/a
Contact e-mail information	The email address the viewers of the data can use to ask questions about the dataset	n/a
Coverage	The coverage area included in the dataset (e.g., Statewide).	Coverage (spatial)
Granularity	The lowest levels of granularity available within the data file (ex. County).	n/a
Define any limitations	Description of any limitations of the Dataset or exclusions.	Rights
URL(s) to additional resources (optional)	URLs to additional resources that may be useful to an end-user	Relation
Narrative Information		
Overview Document	one to two page document that explains the dataset in greater detail, explains the data collection process, and any limitations in the data use	n/a
Data Collection Tool / Data Input	Explanation of the data collection methodology	n/a

Data Dictionary and / or Data file layout	Data dictionary should explain the fields within the dataset in terms of their definition, type, size, and any other pertinent information that describes the dataset	n/a
Benefit of Utilizing Dataset (optional)	Additional supporting documentation can include a data collection/input tool, a benefits document that describes what can be gained from analyzing the data	n/a