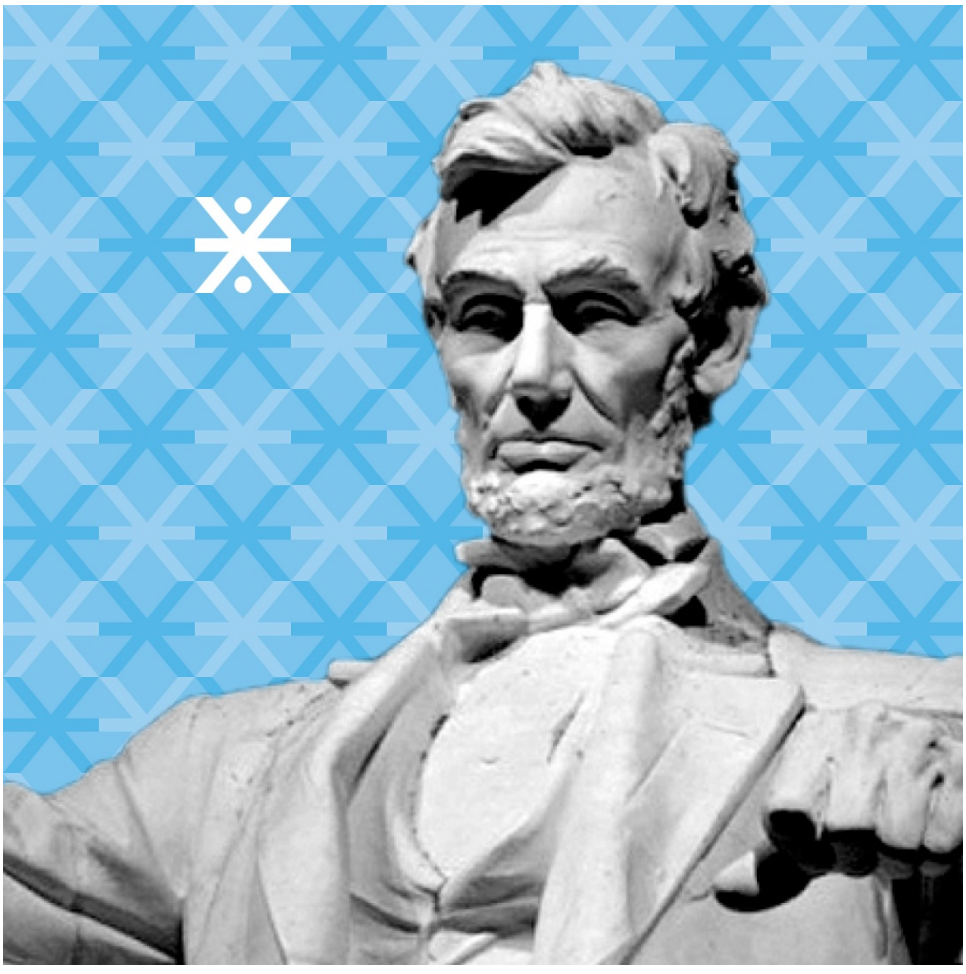
summit

complexity simplified.

Are Your Administrative Data Ready for Public Use?

Association of Public Data Users
2016 Annual Conference
Washington, DC
September 13, 2016



Contact:
J. Anthony Curcio
Principal
202.407.8303 Tel
509.984.8943 Fax
anthony.curcio@summitllc.us

Author:
China Layne, Ph.D.
Manager,
Data Analytics and Research
202.407.8328 Tel
509.984.8943 Fax
china.layne@summitllc.us

Table of Contents

Executive Summary

In recent years, stakeholders have demanded greater access to the data collected by governmental agencies. While there are substantial benefits to publicly releasing administrative data, there are also substantial risks to your organization's reputation. This paper provides an overview of the major considerations your organization should weigh before publicly releasing its administrative data. This paper is organized around four key questions:

- *Why* are you releasing the data?
- *Who* is the audience for the data?
- *What* kinds of data will you release?
- *How* will you release the data?

This paper provides a framework for thinking through these issues. The paper discusses the key decision points for each consideration—and how these decisions may affect future considerations. The paper also discusses standard requirements for the more technical considerations of how to release your administrative data and provides additional resources for in-depth information about these technical considerations. This paper will help guide your organization through the necessary steps that will ensure a successful public release of its administrative data.

Checklist for Public Release of Administrative Data

Key Questions and Considerations	Applicable to Your Agency?	Notes
1. Why are you releasing the data?		
Fulfill Federal or state mandates	◯Y ◯N ◯?	
Pre-empt information requests	◯Y ◯N ◯?	
Provide expected benefits to your organization	◯Y ◯N ◯?	
Provide expected benefits to data users	◯Y ◯N ◯?	
2. Who is the audience for the data?		
Policymakers and practitioners	◯Y ◯N ◯?	
Researchers and analysts	◯Y ◯N ◯?	
Media and general public	◯Y ◯N ◯?	
3. What kinds of data will you release?		
Start-up information	◯Y ◯N ◯?	
Performance management information	◯Y ◯N ◯?	
Cost and expense information	◯Y ◯N ◯?	
Program participation information	◯Y ◯N ◯?	
Audit or monitoring information	◯Y ◯N ◯?	
Tabular data	◯Y ◯N ◯?	
Micro-level data	◯Y ◯N ◯?	
One-time data release	◯Y ◯N ◯?	
Periodically updated data	◯Y ◯N ◯?	
4. How will you release the data?		
Take steps to ensure data quality	◯Y ◯N ◯?	
Create and provide documentation	◯Y ◯N ◯?	
Assess and limit disclosure risk	◯Y ◯N ◯?	
Explore methods for accessing/using the data	◯Y ◯N ◯?	

INTRODUCTION

Increased Public Demand for Government Collected Data

In recent years, researchers, journalists, civil society organizations, and members of the general public have demanded greater access to the data governmental agencies collect as they implement their policies and programs. For its part, the Federal government has become more responsive to these demands for data access.

The Open Government Directive of 2009 requires Federal agencies to implement transparency principles and allows the public to participate in agencies' programmatic activities. The transparency aspect of the directive charges Federal agencies to improve the quality of government information, publish government information online, and set deadlines for implementing these measures.

Similarly, Congress passed and President Obama signed the Digital Accountability and Transparency Act (DATA Act) in 2014. This open government law requires the Department of Treasury and the White House Office of Management and Budget (OMB) to standardize the data on government spending and make it available online. When fully implemented, the DATA Act will make information on all grant- and contract-based government spending easily accessible to the public. In the years following the Open Government Directive and DATA Act, many Federal agencies have begun making more of their data available to the public online.

In addition to these voluntary moves toward open data, various levels of government have responded to calls for accountability from the media and the general public. For instance, in 2013, the *Wall Street Journal* won a lawsuit against the Department of Health and Human Services to gain access to internal Medicare data. The series of articles and searchable database developed from this data won the *Wall Street Journal* an investigative journalism Pulitzer Prize in 2015.

PUBLIC ACCESS TO GOVERNMENT COLLECTED DATA

Department of Labor (DOL) Enforcement Database
DOL provides public access to investigation-related information collected by the department's enforcement agencies, including the Wage and Hour Division (WHD) and the Occupational Safety and Health Administration (OSHA).

Department of Education, National Center for Education Statistics (NCES)
NCES provides public use access to a variety of surveys, such as the Baccalaureate and Beyond Longitudinal Study, and programmatic data collections, such as the Integrated Postsecondary Education Data System.

State of Utah
The state of Utah provides online access to statewide information on a variety of topics such as education, health, and public safety.

Redwood City, CA
Redwood City provides extensive information about residential and commercial development projects at different stages of completion in an interactive GIS-based map.

Similarly, the recent interest across academic and practitioner disciplines in reproducing and verifying published research has spurred governmental research organizations to consider the need for making their data available to the public. Finally, with the rise of data journalism and data science,

analysts and the public are increasingly looking to new sources of data for analysis, such as data collected by the government.

Advantages of Publicly Releasing Administrative Data

Given the current environment, your organization may want to publicly release its administrative data simply to fulfill a governmental mandate or to preempt public demand. However, there are several real advantages to making your organization's data public. First, publicly releasing your organization's data will facilitate secondary analyses by external researchers, which can provide direct benefits to your organization. These secondary analyses can provide insights to help your organization improve its programs and policies. In addition, secondary analyses can uncover innovative ways to use your administrative data to track program performance or measure program impact. Second, the preparation required to release your data and the feedback from data users can help your organization improve how it collects, verifies, and maintains its administrative data, which will increase the overall utility of the data. Third, providing your data for public use can help establish your organization as an authoritative and go-to resource for researchers working in relevant disciplines. Finally, publicly releasing your data will allow other researchers to reproduce research previously conducted by your organization, thereby supporting the growing expectations for research verification.

Main Considerations for Publicly Releasing Administrative Data

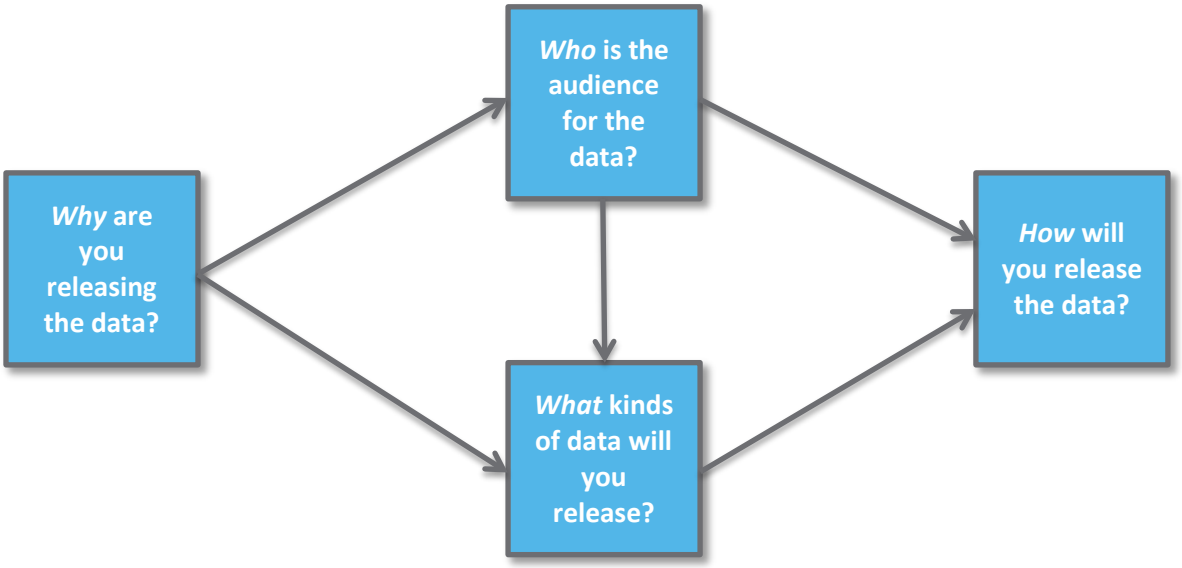
There are many different issues to consider when publicly releasing your organization's administrative data, ranging from basic questions of purpose and audience to more technical considerations of data quality and disclosure risk limitation. This paper will provide an overview of the major considerations and a framework for thinking through these issues.[1] Rather than presenting "the right way to publicly release administrative data," this paper will discuss the key decision points for each consideration as well as the standard requirements for more technical considerations. The paper will also provide additional resources for more in-depth information about some of the technical considerations. The following sections are organized around four key questions your organization should ask before publicly releasing its administrative data:

1. *Why* are you releasing the data?
2. *Who* is the audience for the data?
3. *What* kinds of data will you release?
4. *How* will you release the data?

[1] For additional information on considerations in publicly releasing administrative data, see: M-14-06 Guidance for Providing and Using Administrative Data for Statistical Purposes published by the Office of Management and Budget in 2014, available at https://www.whitehouse.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf.

Four Key Questions to Ask Before Publicly Releasing Administrative Data

The four key questions to ask (and answer) before publicly releasing your organization's administrative data are interrelated. As illustrated in the figure below, the answers given to earlier questions guide the options available for later questions. For example, releasing your data with a specific intended audience, such as the media, will inform *both* what kinds of data your organization releases (e.g. cost and performance information) as well as how your organization releases the data (e.g. on user friendly platforms). In the following sections, we outline the relationships between these four questions and show how together the questions shape your organization's release of its data.



Question 1: Why Are You Releasing the Data?

There are many reasons to publicly release your organization's administrative data, including practical considerations of complying with governmental mandates or media inquiries. However, your organization should consider reasons for releasing the data beyond practical expediency. Specifically, your organization should consider the kinds of benefits the data release may produce *both* for the data-using public *and* for your organization. For instance, does your organization expect the data release to support increased civic engagement or to facilitate general research efforts in the respective practice area? Similarly, does your organization expect to receive direct benefits from the data release, such as data user feedback to improve your data collection efforts or program performance?

By thinking explicitly about the reasons for releasing the data, your organization will be able to better identify the specific audiences for the data and think more strategically about the kinds, amounts, and formats of information to release. For instance, if your organization's primary reason for releasing its administrative data is to fulfill program accountability mandates, the audiences for the data will most likely be the media and general public and the types of data you release will most likely focus on cost and performance information. On the other hand, if your organization hopes to recoup program benefits from the release of its data, the audiences for the data will likely extend to the policymaking and research communities. You will then likely want to provide a much wider range of information in your data release.

Question 2: Who Is the Audience for the Data?

After considering the question of *why* release the data, your organization should ask *who* the target audience will be. There are three major audiences for publicly released administrative data: (1) researchers and analysts, (2) practitioners and policymakers, and (3) the media and general public.

These three types of audiences generally have different interests and levels of skill in working with data. For instance, researchers and analysts are likely to be the most skilled data users and have the widest range of interests for the data. On the other hand, the media and members of the public are likely to have narrower interests for the data (focusing on accountability information) and have fairly limited data skills and experience.

As discussed in the previous section, the intended audiences for your data will be determined in part by the reasons for releasing the data. In turn, by understanding the audiences for the data, your organization can make more informed decisions about the other key questions: the kinds of data to release and how the data will be released.

For instance, if policymakers and practitioners are the primary audiences for the data, your organization may focus the data release on program cost and performance information and release the data so that it can be easily accessed using both standard spreadsheet software and more sophisticated statistical programming packages. Although the media and general public may want the same kinds of program-focused information, these audiences may need tools (such as online tabulation, analysis, or visualization tools) that would not be required by more experienced data users.

Question 3: What Kinds of Data Will You Release?

There are many different types of information that fall under the category of administrative data. Potential sources include program start-up documents (such as grant applications); performance management metrics; and data collected for case management, cost measurement, or program auditing. With careful preparation, nearly all of these types of data can be released to the public. The specific kinds, amounts, and formats of data released will be shaped in large part by your organization's reasons for releasing the data and the intended audiences.

Potential Sources of Administrative Data

- Program start-up documents, such as grant applications
- Performance management metrics
- Data collected for case management, cost measurement, or program auditing

For instance, if the data release is meant to inform the general public or media about your programs, your organization may release basic program cost and performance data in aggregate form (e.g. tables of program outcomes by program sites). However, if the data release is meant to support programmatic research among skilled analysts and practitioners, your organization may release more detailed information at a micro-level (e.g. a dataset of program site and outcome data by program participants).

In turn, by understanding what kinds, amounts, and formats of data to publicly release, your organization can make more informed decisions about *how* to publicly release the administrative data. For instance, releasing micro-level data will require more extensive data documentation, different disclosure limitation efforts, and specific software or platforms for accessing the data than would be

required for releasing data in aggregate form. In addition, releasing data that are periodically updated requires detailed attention to the platform on which the data are hosted and extra care to ensure the comparability and consistency of the data across these periodic updates.

Question 4: How Will You Release the Data?

After considering the questions of *why*, *who*, and *what*, the last issue to consider is *how* your organization will release the data. The specific methods for releasing your data will be determined in large part by the decisions made for the previous three questions. For example, releasing micro-level data for use by researchers and analysts may require more extensive review of the completeness and accuracy of the data than is necessary for releasing tabular data for use by the media and general public. The question of how to release your administrative data involves decisions about several issues, including:

- Data quality
- Documentation
- Disclosure risks and limitations
- Accessibility and usability

How Will You Release the Data?

Data Quality
Organizations should only publicly release administrative data that has the least amount of missing information and the fewest errors possible.

Documentation
Organizations should provide enough information to allow data users to reproduce the analysis results and accurately conduct further analysis.

Disclosure Risk and Limitation
Publicly releasing administrative data raises the risk of exposing the identification of the subject of the data and of being able to attribute private information to an identifiable subject. Organizations should work to identify and limit sources of disclosure risk in the data.

Methods for Accessing and Using the Data
Organizations should strive to release the data in formats and on platforms that reflect the technical capabilities of the intended audiences.

The following sections cover each of these issues in turn and discuss the standard requirements where applicable.

How to Ensure Data Quality

Your organization should be sure to only publicly release high quality administrative data; namely, data that has the smallest amount of missing information and the fewest errors possible. Releasing poor quality administrative data to the public may lead data users to produce misleading or erroneous analysis results. This ultimately compromises the data's utility to the intended audiences. Moreover, releasing poor quality data undermines the organization's reputation, as the public loses trust in the organization's internally produced analyses or other products.

To ensure the quality of the data released, your organization should—at a minimum—review and correct the data for the following issues:[2]

- **Duplicate cases.** To the best extent possible, the data should be released without duplicate cases; only one record in the data should represent one case at the level of the unit of analysis. For instance, if the data provide information on individuals' program participation, one record in the data should represent one participant in the program. If the information in the dataset is updated or revised after collection, the publicly released dataset should only include the most recently updated or revised cases, not both the original and revised cases.
- **Missing values.** The data should have as few cases with missing values as possible and should include values that distinguish cases that are missing information from those cases for which the information does not apply.
- **Invalid values.** In addition, the data should be free of invalid values (e.g. negative values for wage and salary income) and values with incorrect formats (e.g. respondent age in values with decimals).
- **Incompatible value formats**. Values for data elements with known classification or coding schemes (e.g. industry, county and state geographies, race and ethnicity) should be coded in formats that conform to these established classification schemes (e.g. North American Industrial Classification System (NAICS) codes for industry; Federal Information Processing Standards (FIPS) codes for state and county; OMB classification for race and ethnicity groups). This will ensure that the coding for these data elements remains consistent and increases compatibility of the data with external data sources.
- **Outlier values.** Data elements with potentially skewed distributions (e.g. income, hours worked, and weeks worked) should have outlier values identified and resolved through top- or bottom-coding or recoding into intervals.
- **Data inconsistencies.** Finally, any value inconsistencies between data elements with known relationships (e.g. wage and salary income for unemployed respondents) should be identified and resolved through editing.

All of these issues of data quality should be considered before releasing your administrative data, regardless of the reasons for release, the audiences, or the types of data released. For data that will support research and analysis, your organization may want to employ more extensive data quality measures. Many of these data quality issues can be identified and corrected through procedures incorporated into the data collection process itself, rather than just through post-collection review and editing. By employing these methods for reviewing and correcting the data, your organization can ensure that it publicly releases administrative data that meets standard quality requirements.

How to Document the Data

Your organization should provide documentation with enough information to allow data users to reproduce the analysis results and to accurately conduct further analysis. Providing complete documentation with your data will increase the likelihood that the intended audiences will actually use

[2] For additional information on conducting data quality reviews, see: Data Quality Assessment Tool for Administrative Data published by the Bureau of Labor Statistics in 2013, available at http://www.bls.gov/osmr/datatool.pdf.

the data and will help ensure that they use the data correctly. At a minimum, the documentation should include: (a) a complete data dictionary, (b) a data user's note, (c) tables of basic descriptive statistics of the dataset, and (d) an overview of the data collection instrument and process.

DOCUMENTATION: MORE THAN JUST A CODEBOOK

At a minimum, organizations should provide the following data documentation:

- A complete data dictionary
- A data user's note
- Tables of basic descriptive statistics
- An overview of the data collection instrument and process

The data dictionary should include for each variable: its name and label, data type, length, and format, along with a list of valid values and value labels.[3] Additionally, the documentation should include a data user's note with a description of the data structure, identification of unique identifier variables, basic instructions for merging any data contained in multiple tables, and identification of any weights included in the dataset and instructions for weighting and variance estimation if applicable. The information in this user's note will help ensure that data users identify the unit of analysis for the data, understand the interrelationships between any multiple tables of data, and are able to replicate population level estimates if applicable.

The documentation should also include tables of basic descriptive statistics of the sample or population. These statistics will allow users to benchmark key estimates calculated from the data and ensure they are using the data properly. Finally, the documentation should include brief information about the data collection instrument and the sample size, response rate, and methods taken to reduce non-response bias, if this information is applicable to the data source. Taken together, this set of documentation will provide data users with enough information to accurately use the data for research and analysis.

The specific types, amount of detail, and format of documentation your organization provides should depend on the primary audiences for the data and their general skill level. For example, for practitioners, policymakers, and the general public, your organization may need to provide definitions and overviews of general data concepts and avoid using overly technical jargon in the documentation. On the other hand, your organization can provide more detailed, technically oriented documentation for data releases intended for researchers and skilled analysts.

How to Identify and Limit Disclosure Risks

Publicly releasing administrative data introduces the risks of exposing the identification of the subject of the data (e.g. a program participant or organization that is being investigated) and/or being able to attribute private information (e.g. health status or revenue information) to an identifiable subject. These risks are defined as *statistical disclosure*. Releasing administrative data with significant disclosure risks violates several privacy laws, such as the Privacy Act of 1974, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA,) and potentially the Family Education Rights and Privacy Act (FERPA) or the Health Insurance Portability and Accountability Act (HIPAA.) In addition, releasing administrative data that exposes the identity of respondents undermines the organization's reputation as data users grow to mistrust the organization and respondents refuse later data collection requests.

[3] A template data dictionary is included in the Data Quality Assessment Tool referenced in the data quality section on page 8.

Statistical disclosure is a consideration both for data released at the aggregate (tabular) and individual levels (micro-data). Your organization should evaluate its administrative data for statistical disclosure risks and mitigate these areas of risk before publicly releasing the data. There are minimum standards and generally recognized methods for assessing and mitigating statistical disclosure risk in both tabular and micro-level data.[4] For tabular data, organizations first identify individual cells in the table that pose a statistical disclosure risk, generally by using methods such as the (n, k) rule,[5] and then mitigate the disclosure risk using, at a minimum, primary and complementary suppression techniques.[6]

For micro-level data, organizations first ensure that no personally identifiable information (e.g. names, full addresses, social security numbers, etc.) are included in the dataset. Next, organizations identify individual values of data elements that pose a statistical disclosure risk, using, for instance, a minimum threshold for the necessary number of respondents with each value of the data element. Once the data element values that pose a disclosure risk are identified, organizations mitigate this risk using a variety of techniques including data blurring (least disruptive), data perturbation, or data suppression (most disruptive).

Statistical disclosure limitation can be a complicated process, but there are many resources available to guide and assist your organization. Attending to the requirements of statistical disclosure is vital to both protect the privacy of the subjects of your administrative data and to ensure that your data are used properly once released to the public.

How to Ensure Data Are Accessible and Usable

Your organization should be sure to publicly release the administrative data in formats and on platforms that reflect your organization's ongoing data release expectations and that correspond to the technical capabilities of the intended audiences. For instance, if your organization expects that the data will be updated regularly (e.g. program participation information collected and released over several years), you may consider flexibility and reliability as important factors when deciding which data formats and platforms to publicly release the administrative data. These considerations could lead your organization to release the data in comma separated values (CSV) files on a file transfer protocol (FTP) site, which provides access to the data in a format that works well with many types of analysis software and is on a platform that is stable, secure, and allows for additional datasets to be easily added.

In addition, if your intended audiences are primarily non-technical, you may not want to release the data in highly technical formats such as Application Programming Interfaces (API) or JavaScript Object Notation (JSON), which would be difficult for non-programmer data users to employ. Similarly, your organization may consider including on-site tools (e.g. table generating or graphing/visualization tools)

---

[4] For additional information on statistical disclosure, see: Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology published by the Federal Committee on Statistical Methodology, available at https://fcsm.sites.usa.gov/files/2014/04/SPWP22_rev.pdf.

[5] The (n, k) rule proposes that a table cell is at risk of disclosure if a small number (n or fewer) of respondents contribute a large percentage (k percent or more) of the total cell value.

[6] Primary suppression involves not publishing individual cells that have identified disclosure risk. Complementary suppression involves not publishing other cells in the table if that information can be used to accurately estimate the value of the primary suppressed cell.

to increase the utility of the administrative data for non-technical audiences. Ultimately, specifics about the intended audiences, purposes, and kinds of administrative data to be released taken together with your organization's technical capabilities will determine what formats and platforms your organization uses to provide access to and use of the data.

Conclusion

In recent years, stakeholders have demanded greater access to the data collected by governmental agencies. This paper provided an overview of the four major questions your organization should consider before publicly releasing its administrative data: (1) *why* are you releasing the data, (2) *who* is the audience for the data, (3) *what* kinds of data will you release, and (4) *how* will you release the data. This paper provided a framework for thinking through these issues. The paper discussed the key decision points for each question and showed how later questions are dependent on decisions from earlier questions. The paper also discussed the standard requirements for the more technical considerations and provided additional resources for in-depth information about some of the technical considerations. This paper will help guide your organization through the necessary steps to ensure a successful public release of its administrative data.

About Summit Consulting, LLC

Summit is a specialized analytics advisory firm that guides Federal agencies, financial institutions, and litigators as they decode their most complex analytical challenges. Summit's staff of economists, econometricians, and research scientists use quantitative techniques to assist our clients as they model risk, evaluate program performance, and predict future performance.

At Summit, we solve complex analytical challenges with unparalleled customer service and extensive client collaboration. The solutions are complete only when they are understood by our clients and solve their problems. Our distinct capabilities include program evaluation, applied statistics and economics, mortgage finance, financial services, Federal Credit modeling and forecasting, and litigation analytics.

Summit hosts a solutions-focused academic environment and is dedicated to staying at—and pushing—the forefront of analytics best practices. To that end, our staff members present research at conferences worldwide and regularly partake in intensive in-house technical trainings. Our principals, academics, and research scientists are recognized experts in their fields, and they are capable of leading large and small solution teams.

Summit: complexity simplified

www.summitllc.us