2023

City of Madison
Data Management Guide

HOW TO CREATE, MAINTAIN, AND REPORT DATA

DATA GOVERNANCE TEAM

CITY OF MADISON | 210 Martin Luther King Jr. Blvd., Madison, WI

Authors, Contributors, and Sponsors

This guide would not have been possible without the contributions of the following:

Jessica Jones, Finance, Data and Innovation Team & Data Governance Team
Karalyn Kratowicz, Human Resources, Data Management Working Group & Data Governance Team
Eleanor Anderson, Finance, Data and Innovation Team & Data Governance Team
Milena Bernardinello, Planning, Data Management Working Group
Thomas Dull, Police, Data Management Working Group
Patrick Empey, Planning, Data Management Working Group
David Faust, Information Technology, Data Management Working Group & Data Governance Team
Suzanne Fichtel, Police, Data Management Working Group
Candice Kasprzak, Engineering, Data Management Working Group
Stephanie Mabrey, Finance, Budget Team
Mary Morris, Finance, Data and Innovation Team
Julia Olsen, Public Health Madison & Dane County, Data Management Working Group & Data Governance Team
Katy Petershack, Community Development, Data Management Working Group
Adam Pfost, DPCED Office of the Director, Data Management Working Group
Sangeetha Shreedaran, Finance, Data and Innovation Team
David Singer, Finance, Data and Innovation Team
Karl van Lith, Human Resources, Data Management Working Group
Bradley Wollmann, Human Resources, Data Management Working Group & Data Governance Team
Iliana Wood, Community Development, Data Management Working Group

The guidance of the following:

Roger Allen, Attorney's Office, Data Governance Team
Emily Clavette, DPCED Office of the Director, Data Governance Team
Norman Davis, Human Resources, Data Governance Team
Sarah Edgerton, Information Technology, Data Governance Team
Katarina Grande, Public Health Madison & Dane County, Data Governance Team
Christine Koh, Finance, Data Governance Team
Simone Munson, Police, Data Governance Team
John Patterson, Police, Data Governance Team
Adriana Peguero, Attorney's Office, Data Governance Team
David Schmiedicke, Finance, Data Governance Team
Nate Shipley, Finance, Data and Innovation Team & Data Governance Team

And the feedback of forty data users across twenty-seven City agencies, including the City Data Stewards for these agencies.

Contents

Introduction

Data, data, data. It is something you hear about all the time. You know you need to use it to provide City services, and you know you need a way to manage it, but where do you start?

Our City's goal is to make data-informed, equitable decisions. To do that, we need relevant and quality data. This guide will go through the considerations needed to create, maintain, and report on data at every step of its lifecycle. You may be involved with only some of these steps, but with each person following these best practices at every stage, the City will benefit.

Glossary of Key Terms
You will noticed that some words in this introduction have been **bolded and colored in light blue**. It indicates they are important concepts, and their summarized definitions along with quick examples are available in the Guide's Glossary of Key Terms.

Foundational Concepts

Data governance refers to the means by which an organization makes decisions about its information assets. It systematically establishes and enforces policies, procedures, roles, and responsibilities for the collection, maintenance, and use of data with the intent to organize program staff to collaboratively and continuously improve data quality throughout the organization.

The City of Madison seeks to establish a common understanding of data management practices across the organization. **Data management** refers to the implementation of practices to ensure the overall management of the availability, usability, integrity, and transparency of data across the City of Madison. The difference between data governance and data management is the difference between oversight and execution of policies, procedures, roles, and responsibilities.

Data analytics refers to the use of data analysis methods to describe, predict, and improve organizational performance or solve problems. The difference between data management and data analytics is the execution of policies to ensure the availability of quality information and the use of data itself.

Fundamental to data management and analytics is the data itself. *But what is data?* **Data** is defined as information, especially facts or numbers, which can be examined, considered, and used to inform decisions. You may often hear data in the context of the term **dataset**, which is a collection of data that is related by content and structure. Another common data term is **database**, which is a systematic collection of data that is stored to facilitate its access, modification, and deletion in conjunction with various data-processing operations. A **data point**, meanwhile, is a single piece of information. The following table provides City examples.

City of Madison Example	
Data	The year to date actual expenses in MUNIS for a given agency
Dataset	The year to date actual expenses for all agencies
Database	The storage system for all data stored in MUNIS
Data point	A single expenditure recorded in MUNIS

Data is one of the most valuable assets of the City. Data provides information that allows us to inform decisions, follow past trends, monitor our work, and estimate future consequences. In understanding and using data effectively, the City is able to most efficiently and effectively provide services, while also demonstrating its commitment to the transparent use of public resources.

Remember, everyone in the organization must take ownership of their data. As long as issues with data are seen as an IT problem, it will be difficult to make improvements to data quality.

Sidebar 1: *Data vs. Records*

You may be familiar with records because the City has a robust records management program due to Wisconsin's strong Public Records Law. Like data, **records** are also defined in terms of information: they are materials created or kept by an organization, containing information relating to the function of that organization (Wisconsin Public Records Law Compliance Guide).

So what is the difference between data and records? Technically speaking, there is no difference: as you can see by comparing the definitions, data is records, and records are data, because both are information. However, the terms are often used in different contexts.

People often use "data" to mean one or both of the following:

- Information in a structured format.

  Example: a table with rows and columns

- Information considered in aggregate, that is, many data points together. Typically when people talk about data, they are less interested in what any one data point (or record) says and more interested in what patterns the dataset considered as a whole can reveal.

  Example: looking at a single water station's pumping output would say nothing about total water use in the city, but looking at the output of all water pumping stations over a year would.

People often use "records" to mean one or both of the following:

- Information in an unstructured format

  Example: a Word document or PDF

- Information considered individually or in small amounts, such as the records of a single employee or investigation. Typically, when people talk about records, they are interested in taking action or making decisions based on one or a few records and are uninterested in the rest.

  Example: an employee may be disciplined based on records pertaining to them, but the records of all employees would not be relevant to that process.

However, information typically considered to be data can be used as records, and information typically considered to be records can be used as data. For example, someone analyzing trends in employee pay may focus on details about their own pay as a familiar calibration point, while advanced data analysis techniques like machine learning can use unstructured information as their data. It all depends on context.

This guide uses the term "data" because it is ultimately interested in helping people use aggregate information to make decisions. However, this guide begins with the process of data creation and collection. Sometimes this information is created in small or unstructured batches, and **depending on your context, what this guide calls data in that stage may be what you think of as records.**

Purpose and Use of the Data Guide

This guide is intended to give City employees who work with data in any capacity a baseline knowledge of best data practices, including ways to eliminate bias in the process for more equitable data. Its ultimate goal is to advance data-informed decision-making and support Citywide data initiatives, such as **Results Madison**, a component of the City's strategic framework that uses data-based indicators to help us better understand our services and where to target improvements. Neither this guide nor any associated trainings are intended as training on specific software programs.

A critical question we need to ask ourselves in implementing Results Madison, and in working with data in general, is: *how do we get the right data to tell the story of our City and help us understand our services?* Having the best management practices for creating, maintaining, and reporting data will ensure the most accurate and complete picture of our city for agencies, researchers, and community members alike. *And so this guide is divided into three main sections:* data creation (*Phase 1: Create*), data maintenance (*Phase 2: Maintain*), and data reporting (*Phase 3: Report*). In these sections, you will explore your way through the data lifecycle to build a shared understanding of data management. In the appendices, you will find discussions of crucial considerations that embrace the whole data management process, the City of Madison Data Standards, and some resources, such as a project charter template.

A Data Journey
In each section, look for the "A Data Journey" pages to follow the story of a fictitious team going through a data journey.

As actors of change, **City staff who work with data at any level are expected to engage with the Data APM and Guide**. These materials' standards and practices are intended to be broad and general enough to cover many data contexts observed in the City, and data users are expected to refer to them and conduct data-related work in accordance with them.

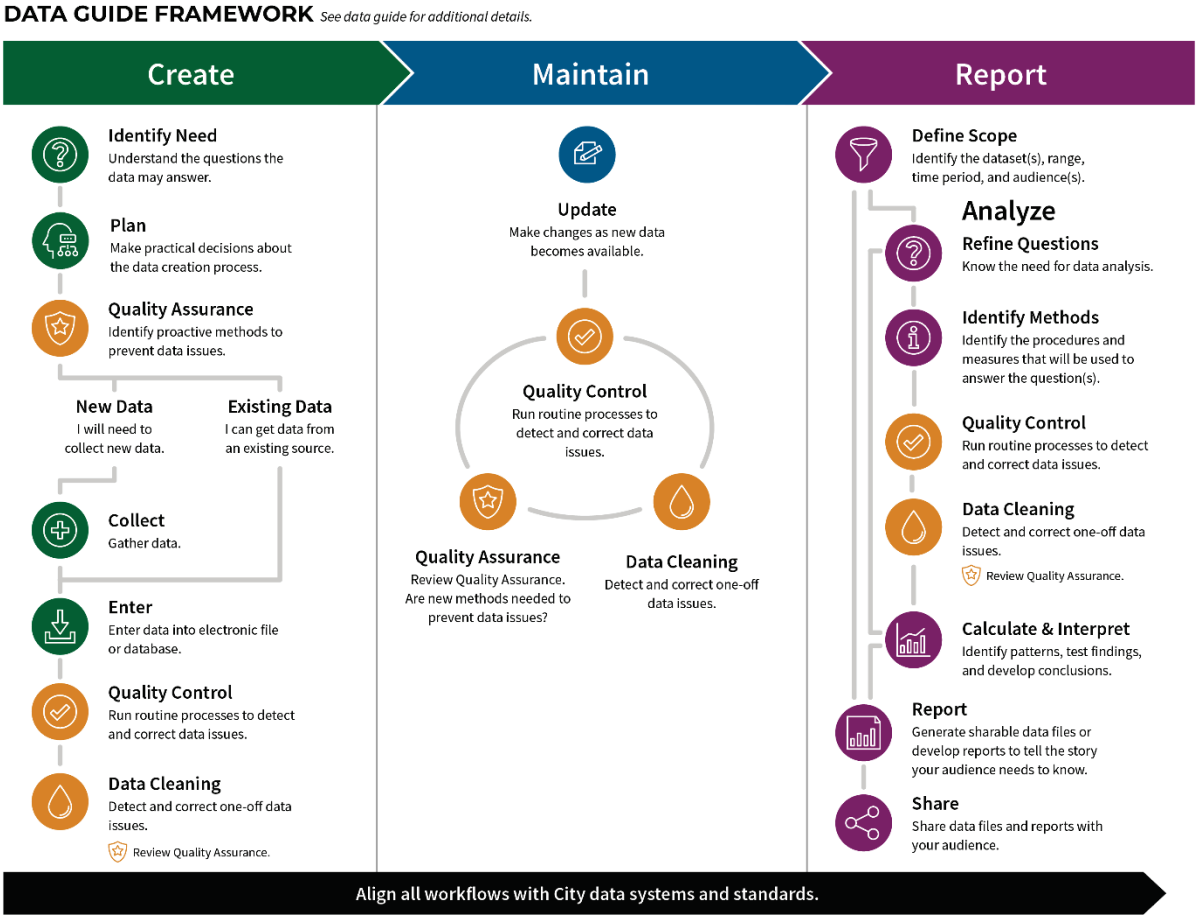
However, it is acknowledged that these standards and practices will not fit the needs of every agency and every situation. This may be because they do not cover a specific subject area, because of external factors like mandatory reporting requirements or vendor capabilities, or other situations. In these cases, when data users face difficulty implementing these standards and practices, data users are expected to attempt to find solutions that are in alignment with both the APM and Guide, and the external requirements; and to use their best judgment to tailor the APM and Guide's standards and practices to their situations.

Moreover, in these cases, data users with needs not met by the existing APM and Guide should not simply ignore these standards and practices. This leads to the very problems with data quality and siloing the APM and Guide are intended to address, and ultimately hamper the City's ability to make data-informed decisions. Instead, data users are expected to engage with the Data Stewardship Program and the Data and Innovation Team to update these materials to reflect their needs. The APM and Guide are living documents, and in this way, we will continue to build together comprehensive data standards and practices to meet the needs of our City.

Data Management Framework

The heart of this guide is the Data Management Framework, which is a high-level overview of the data management process and the steps involved in each of the different phases. The process, at its basic level, consists of raw data being created (i.e., collected), maintained, and reported or transformed into actionable insights through analysis.

The image below is a flowchart that summarizes all the steps in the framework:



Please see Appendix E: Printable Data Management Framework for a larger version that you can print.

Data Management Framework Users

No matter your role in the organization, you are likely involved in one or more of the three phases of the data management process. That is because as more data becomes available, nearly everyone in the organization will inevitably touch data. In other words, almost everyone in the organization is considered a Data Management Framework user.

Data Management Framework Users: all city employees who interact with data at any level.

- Those who collect data throughout the course of their duties as a record of their activities.
  For example, by entering the time they began an activity.
- Those who collect and aggregate data as a major part of their job.
  For example, by entering information from resident forms into a spreadsheet.
- Those who perform data analysis and report findings.
  For example, by creating a bar graph to show the most common reasons for room reservations.
- Those who use data as a tool to aid in performing their responsibilities and inform decisions.
  For example, by using the most common reasons for room reservations graph to decide how to allocate resources.

Stakeholders and Collaboration

Data management is a collaborative effort across the organization. It means that regardless of the type or level of your involvement in the process, you will work with multiple partners and stakeholders, who will also have various types and levels of involvement. The Data Management Framework assumes the critical importance of collaboration.

What form this takes will be different for each project, so this guide does not explicitly mention collaboration in its discussion of each step. However, as you read, think about how you can collaborate and ensure the appropriate participation of diverse stakeholders.

Here are examples of ways you may collaborate with stakeholders:

- Consulting stakeholders from diverse backgrounds from the beginning to define the vision and plan for the project.
  For example, conducting a kick-off meeting to gather stakeholders' data needs as part of the Identify Need step.
- Scheduling regular meetings and/or other forms of communication with stakeholders to share milestones and receive feedback on partial deliverables.
  For example, having monthly meetings with a project sponsor and quarterly meetings with an external organization involved in the project.
- Reaching out to appropriate subject-matter experts to better understand the context or structures related to the data.
  For example, asking questions to an epidemiologist to find out the social and public health context of data related to a certain infectious disease.
- Reaching out to residents to learn about their lived experiences and contextualize the data.
  For example, partnering with community-based organizations to hold community meetings and/or resident panels.
- Involving stakeholders in the process of understanding the story being presented by the data.
  For example, presenting findings to appropriate stakeholders and asking for their interpretation of the data and analyses.
- Ensuring that all stakeholders understand the need for data collection, maintenance, analysis, and/or reporting.
  For example, including notes in data collection methods, such as surveys, and/or reporting methods, such as written reports, about why specific data is being collected and/or analyzed).

Sequence of Steps and Iterations

Lastly, note that the framework describes an almost linear journey. That is because it is easier to establish and visualize processes that follow a linear sequence, and some steps work as foundations for the following steps.

However, individual aspects of each data project may result in a sequence break, such as skipping steps (e.g., if your data creation work involves existing data, you may skip the Collection step) or returning to a previous step (e.g., your findings in the Calculate & Interpret step may prompt new questions, or you may have to toggle between Entry and Data Cleaning several times in order to get the most accurate data). Additionally, some steps may occur concurrently.

The order of steps outlined in the framework aims to align industry best practices with the most common breaks and cycles observed at the City of Madison. Still, you may adjust your data journey differently according to your needs. But keep in mind that each phase and step in the framework is essential to tell the story of how much the City does, how well the City does it, and whether anyone is better off as a result.

Overarching Considerations

As you read this data guide, there are a few considerations that you should have in mind throughout.

Data Ethics

Data is a powerful tool you can use to improve residents' lives. However, it needs to be used responsibly and with intention. **Data ethics** is a code of behaviors and practices that helps ensure the City of Madison handles and uses data ethically.

In this guide, you will learn to keep the resident in mind when working with data. This means addressing equity and privacy concerns, as well as committing to only collecting and using the data you need. This also means considering the scale and impact of the data's end use and being transparent about it.

For instance, some questions that may help you understand the scale and impact of your work are:

Will the data be used to establish eligibility for programs and benefits?
Will it be used to establish fees?
Who will be impacted by its use?
How should you communicate the objectives?
Which procedures and expectations should you establish to ensure transparency between data users and curators?

In essence, the power that comes from data must be used with care. You should always take into consideration principles of fairness, privacy, transparency, and accountability (Cognizant).

Data Equity

By providing information and insights, data has the power to shape conversations about policies, programs, and resource allocations. It can help decision-makers make fairer decisions by providing an understanding of the lived experience of residents and the identification of invisible disparities and their root causes. For example, the West Oakland Environmental Indicators Project, a resident-led, community-based organization, used participatory research methods to give a voice to residents and use their data to fuel environmental justice legislation and regulatory policy in the City of Oakland and the State of California.

Unfortunately, not all data helps us make fairer decisions. Many people believe that "numbers don't lie," and that data is inherently neutral. But this is not the case**. Data is collected, analyzed, interpreted, and distributed by people, who bring to their work their subjective experiences and potential biases, even unintentionally** (Hawaii Data Collaborative). As a result, one of the most significantissues in working with data today is that people create products, analyses, and research without purposefully thinking about equity in every step of their work (We All Count Project).

Equity is a crucial aspect of fair decision-making and one of our priorities as a City. It refers to the allocation of resources and opportunities to provide equal outcomes to all residents. It is based on the recognition that each resident has different needs and that social issues, such as racism and sexism, impact some more than others (United Way NCA). The City's Racial Equity & Social Justice Initiative (RESJI) helps agencies include racial equity and social justice in their decisions, policies, and services.

To produce and use data that can support equity in decision-making, you will need to ensure that your data work is also equitable. In other words, you will have to apply data equity concepts in each step of the data management process.

Data equity refers to the ways we ensure fair and inclusive data work. It takes into account how data can make social injustices better or worse. It involves:

- Understanding how your data practices can impact groups of people who face social injustices due to discrimination, especially on the grounds of race, gender identity, sexual orientation, age, physical ability, language, and/or immigration status.
- Identifying your own biases and how they can affect your data work.
- Ensuring that your data practices are not creating or reinforcing social injustices, even unintentionally.
- Exploring ways you can use data to increase equity and fairness for those who are in social disadvantage.
- Ensuring that your data products and practices are transparent, understandable, and accessible, while respecting privacy limitations.

As you will learn in this guide, every step of the data management framework involves human design, human judgment, and human action. Thus, there are many opportunities for our biases to enter the data.

Equity Considerations

In each step, look for the "Equity Considerations" page to learn how to build an equitable data practice.

The following example illustrates how biases can create or reinforce iniquities in different steps of the framework.

Example: Bias and Data Inequity

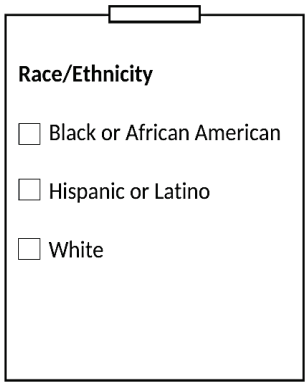
The *City of Example* wants to create a program to address racial/ethnic disparities in the outcomes of *a certain health issue*. In order to understand these disparities, the City started collecting racial/ethnic data for the *health issue* instances and used it to prepare a report.

The racial/ethnic profile of the *City of Example*, from largest to smallest group, is 60% White, 20% Black or African American, 10% Hispanic or Latino, 8% American Indian or Alaskan Native, and 2% Other Races/Ethnicities.

Bias in Collecting Data

In order to collect racial/ethnic data, John, an employee of the *City of Example*, created the form to the right. Unfortunately, instead of checking the City's demographic profile, John listed racial and ethnic groups based on his own experience (i.e., based on what he sees around him).

Equity issues: Exclusion of racial/ethnic groups that are often excluded due to being minorities.



Race_Ethnicity
Hispanic
Black
White
White
Hispanic
Black
White
White

Bias in Entering Data

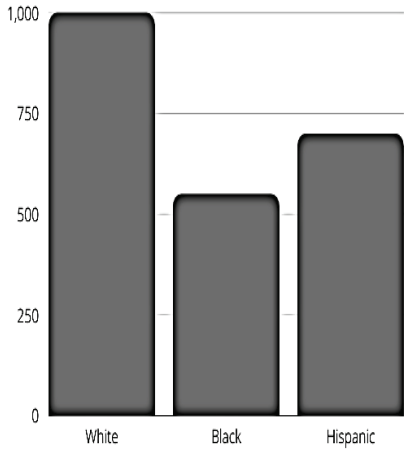
After the collection step, Elizabeth, another employee of the *City of Example*, entered the race/ethnicity data in a spreadsheet column, as demonstrated in the image to the left. Elizabeth removed the racial identities after "or" based on the wrong assumption that they are just alternative names for the same identity. Moreover, she did not document this change.

Equity issues: If the reporting team follows Elizabeth's labels, people who identify with the removed racial identities will not feel represented.

Bias in Reporting Data

Finally, another group of City employees built a report using the collected data. The graph to the right was created using Elizabeth's incomplete labels, and the team chose to use the count of health issue incidents by racial group to show disparities.

Equity issues: this graph excludes racial groups, reinforces whiteness as the norm (white listed first for no specific reason), and incorrectly communicates that white residents are at higher risk (using the count of incidents is not a good measure for disparities because the City's racial profile is not evenly distributed).



Data Privacy

Privacy is a top concern among city employees and residents. When you gather data, the public is putting their trust in you to use that information respectfully and responsibly. Throughout this guide, you will be asked to consider the following questions:

Necessity
• Do you need to ask this question? • Do you need to collect this information?
Openness
• Do the people who will be impacted by your data work understand why you are collecting data? • Do they know how the data will be used?
Respect
• Are efforts to collect desired information reasonable and respectfully requested?
Confidentiality
• How is this information being stored and collected? • Who can have access to it?

In addition to considering these questions, you also should get familiar with the local, state, and federal laws that protect privacy.

The following table presents three broad **data privacy** categories and their implications:

Category	Description	Implications
Public	Data that can be publicly disseminated without any concerns.	When *public data* is made available easily, such as through the City's Open Data Portal, it reduces the need for different groups collecting the same information. This is beneficial for data practitioners and for the population, especially those who are usually targeted for data studies.
Protected	Data that is protected by law or regulation and can only be shared or accessed by a limited group or through a limiting procedure; if cleaned to remove certain information, or aggregated, this data could potentially be shared.	When *protected data* is made available publicly, it puts the people represented by the data at risk. They could become subject of identity theft, stalking, harassment, or other dangers.
Sensitive	Data that is not regulated like *protected* data, but in its raw form poses security concerns and could potentially target individuals or pose other concerns; if cleaned to remove certain information, or aggregated, this data could possibly be shared.	When *sensitive data* is made available without measures to remove its "sensitive" part, it can also put the people represented by the data at risk. Like protected data, people could become subject of identity theft, stalking, harassment, or other dangers.

Data Silos

A **data silo** is a collection of information in an organization that is isolated from and not accessible by other parts of the organization (Alooma). Even though we are all part of the City, our agencies tend to operate in silos, which are usually a result of security, systems, and policy constraints within the agency. However, some silos can also have cultural roots. As you read this data guide, think about how you can help break down data silos across the City.

In order to best utilize the information we collect, data management best practices ask us to consider sharing data while following legal privacy requirements, such as protecting personally identifiable information.

The Negative Impact of Data Silos…	
On the City Agencies	• Data may be inconsistent between agencies and/or workgroups. • Different methodologies across agencies or even staff within a single agency may produce different analysis results. • There may be no single source of truth to reference as multiple agencies may work to analyze the same data, potentially producing different results. • Redundancy in workload/analysis. • Data may be incomplete and therefore only tell a partial story.
On Public Perception	• Data is not useful. • May draw inaccurate conclusions because they only have part of the story. • Answer may vary between agencies. • Mistrust in government.

As we work to break down data silos, **dataset inventories** will become useful tools for agencies to understand what data other agencies collect. The dataset inventories are agency-based documentation of the data assets of each agency and the associated metadata, including primary points of contact for each dataset. The summarized copy of the combined dataset inventories is available to the public on the open data portal, and each agency's full version is available internally.

Data Ownership and Stewardship

Data ownership refers to both the possession of and responsibility for information. This includes the ability to access, create, modify, package, or remove data, the right to assign these access privileges to others (Department of Health and Human Services), and the responsibility to make use of these abilities in a safe and secure way. It also includes the responsibility to maintain accurate, updated, and usable data.

Everyone who interacts with a dataset is responsible for it at some level, however, the data owner has the ultimate responsibility for the data quality, accuracy, and availability. This promotes both accountability and data quality.

One point of caution: Sometimes, data owners get so caught up in the protective aspects of their duties that they create data silos, where they keep data to themselves and do not share it with other parts of the organization**. However, part of a data owner's responsibility is to make sure data is shared where appropriate.**

Data stewardship refers to the implementation of data governance policies and the oversight of data management practices within departments. At the City, through the Data Stewardship Program, each agency has a designated data steward who works to ensure their agency manages and uses data in accordance with City-wide policies, standards, and best practices.

Some examples of data stewards' actions are:

- Implementing the data governance policies laid out in the Data APM, such as data retention.
- Coordinating tasks outlined by data governance policies, such as the update of dataset inventories.
- Becoming familiar with the Data Guide recommendations and identifying opportunities to implement them.
- Communicating to staff within their department about the importance of following data work best practices and encouraging them to follow the Data APM and Guide recommendations whenever possible.
- Supporting and helping staff within their department to implement changes and clarify questions related to the Data APM and Guide.
- Working cooperatively with other data stewards and agencies to improve data interoperability and portability.
- Serving as a point of contact for internal and external inquiries relating to data at the agency, such as helping staff from another agency access needed data within the steward's agency.
- Participating in update iterations and providing feedback to improve the Data Guide, considering their own department needs and the (sometimes conflicting) diversity of departmental needs.
- Reaching out to request additional support or communicate an urgent need for ad hoc updates.
- Contributing to other data stewardship activities around supporting initiatives, such as coordinating and ensuring all staff receive appropriate data training.

Sidebar 2: *Data Steward vs. Records Custodian*

Data stewards and **records custodians** deal with similar subject matter and perform similar actions to ensure the quality of data and records. The main difference between these two titles is:

Data stewards work to improve the handling of *aggregate* data and systems.
For example, changing how we record information about gender of all clients.

Records custodians often work with the *content of individual* records.
For example, finding records for an open records request related to a specific topic.

An employee can be both a data steward and a records custodian for their agency, since both roles require familiarity with, knowledge of, and access to the agency's data and records.

This page left intentionally blank

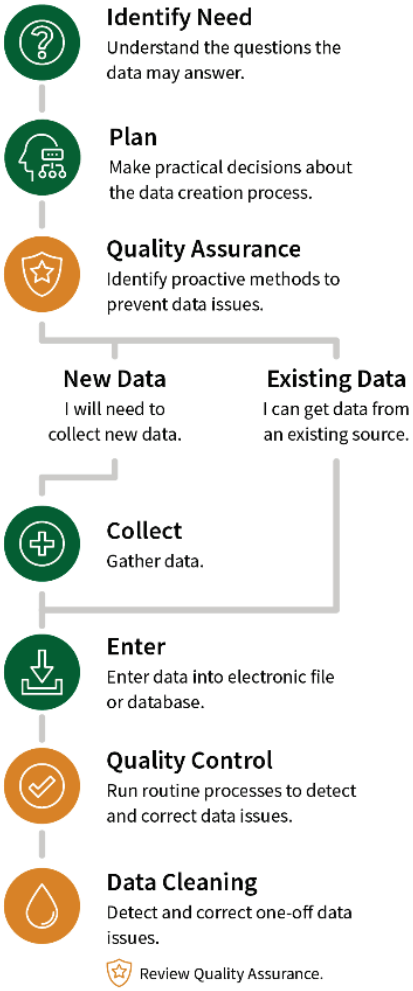
Phase 1: Create

"Create" does not mean making up data. Rather, the Create phase is about laying a solid groundwork before and during the initial round of data gathering to ensure data quality now and in the future.

The Create phase has a total of seven steps, as presented in the image to the right. First, you will begin by identifying the need for data, followed closely by planning the technical aspects of its creation. Then, you will establish practices to ensure the quality of the data that will be created. Next, if you are not working with existing data, you will move on to collecting new data. After acquiring either new or existing data, you will enter the data into a computer system, such as a database. Lastly, as the final steps, you will run routine and one-off processes to detect and correct data issues.

Notice that it is fairly common to skip the Collect step when working entirely with existing data. However, you may also have to return to previous steps or work concurrently in two or more steps according to your individual data needs and work context.

As you work through the steps of this phase, you will likely benefit from collaborating with stakeholders, so please also consider the best ways to do that.



Identify Need

Before you begin creating data, you must think carefully about *why* you need to create it and *what* questions it may answer. This step is crucial because the choices you make here will affect your entire project and any future projects that could use your data.

So, why do you need to create data? Well, numerous projects and policies may call for data. Indeed, it may be required by the federal or state government. Additionally, internal and external stakeholders, including other city agencies, residents, and researchers, may request data to help understand city services. Needs may also be identified based on project-specific questions. And they can also be based on anticipation of future projects; that is, you may not know now exactly how you will use the data, but it will most likely be used in a future project. Once you understand the "whys" behind data creation, proceed to understand the "whats."

What questions should the data answer? This is an important question because it will help you define the right data to create, improving the quality and efficiency of your work. For example, if you create data that will not be used to answer current or future questions, you are wasting resources and slowing operations. Similarly, if you do not create the right data to answer current or future questions, you will either be restricted to less comprehensive analyses or spend more resources to recreate the missing data.

Equity Considerations: Identify Need

As the first step in the Data Management Framework, Identify Need is the foundation for all the following steps. So, it is crucial to define how you will ensure that your data work will not be biased and unfair. Some actions that you should take in this step are listed below:

Be aware of your own biases. Identify any over-generalized beliefs you may have about a particular group of people. Harvard's Project Implicit is a tool that can help you with this.

Make sure you understand the social factors related to your data work. Research how events, policies, and other social factors affect the data you will be creating.

Identify the groups of people who are unfairly affected by data practices. Take the City of South Bend, Indiana, as an example. It developed an urban renewal project driven by data on abandoned houses. But its data work did not include identifying the profile of who owned the houses chosen for demolition. As a result, the City failed to predict the disproportional impact of the project on Black and Latino communities. A similar local example is Madison's urban renewal practices that target the Greenbush neighborhood in the 60's.

Assess how your data work will impact residents, especially those identified above. Consider how your objectives, questions, and methods may affect groups and individuals. More specifically, identify the benefits and risks for those affected. Even if you identify that your work will mostly bring great benefits, check if those will outweigh the risks for all groups. Sometimes, the people who carry the biggest risks may not be the ones who benefit.

Identify the data that will not be collected and its impact on residents, especially those identified above. What you choose not to collect data on is often as important and impactful as what you choose to collect. In the case of the South Bend example previously presented, the City did not collect data about what residents would prefer or about the race/ethnicity of those affected. These two choices resulted on disproportional impact on Black and Latino communities. However, in some situations, you may purposefully choose not to collect certain data to protect residents and ensure equitable outcomes.

Involve the community to identify the right needs – when reasonable and possible. You can reach out to those who will be affected by your work to learn what data they think is relevant to improve their lives. For example, you can hold listening sessions, attend public hearings, and engage local nonprofits.

Be aware of sensitive topics and how they affect groups and individuals. Identify data questions that can make people feel unsafe and uncomfortable. For example, social identity is sensitive data because of its link to discrimination. So, make sure you only collect sensitive data when truly necessary. If you must collect it, use appropriate language and explain how you plan to use it.

Some additional questions that might need to be asked prior to creating data include:

- Who needs this data? Why do they need it? How will they use it?
- Which people, voices, and stakeholders should be part of the project, and in what capacity? These could be community members, staff from other agencies, or others.
- Who will potentially use this data in the future? What are all possible additional questions they may ask?
- Will what you collect today make sense or be useful to others who will later analyze the data?
- Will those who are collecting the data know why they are collecting it and how to appropriately collect it?
- Will the public understand what data is being collected, why it is being collected, and how it will be used?
- How can you balance deciding which data needs to be created with maintaining operational efficiency? Will your data ever be used, or are you wasting resources?
- Do existing enterprise software systems or databases already contain the data you need, or will you need to collect the data from scratch?
- If your data creation need involves updating existing datasets, should old values of data be kept or overwritten?

Examples of Data Creation Needs

Understanding City Services

Metro Transit creates and maintains data on ridership by bus stop to better understand its public transportation services. A need to understand City services framed this.

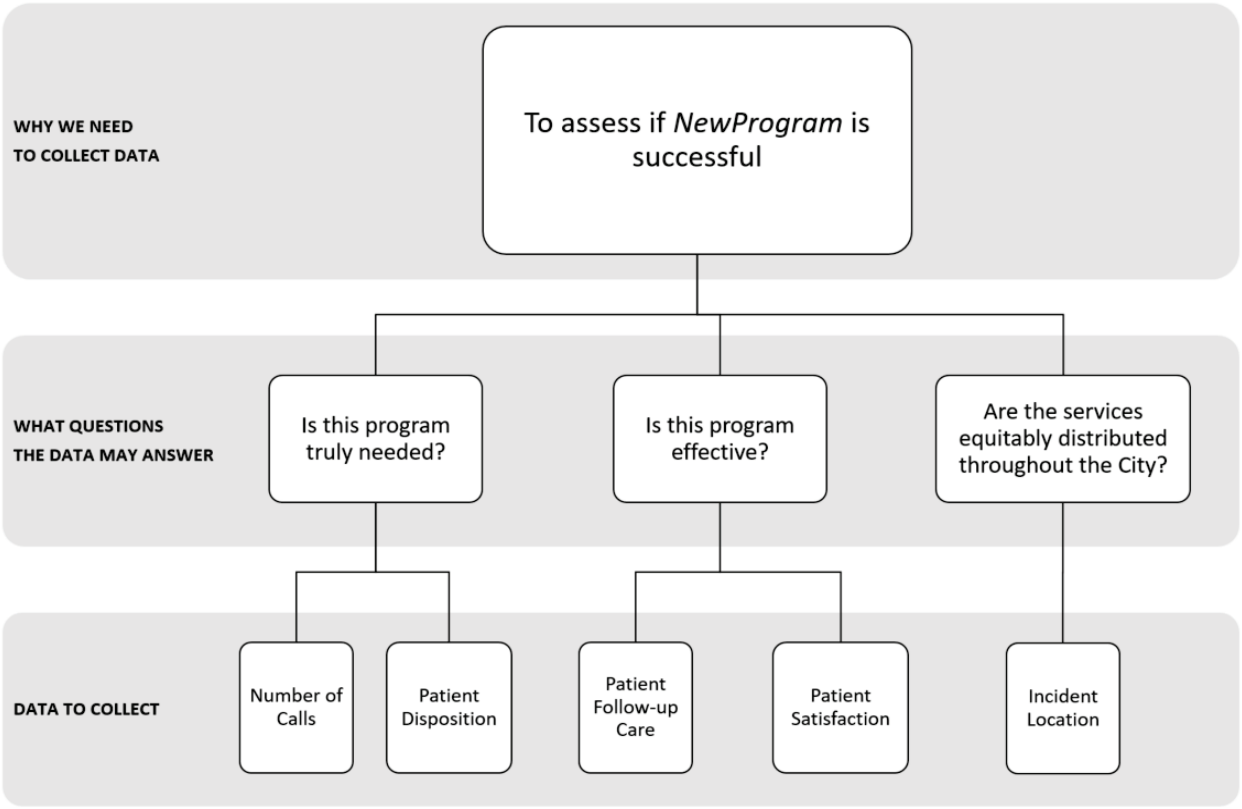
Answering Project-Specific Question

Specific Common Council questions led to a study of the density and impact of alcohol outlets around the City. A project-specific question framed this.

A Data Journey: Identify Need

The *City of Example* is planning to launch a new program within the Department of Public Health. To evaluate *NewProgram*'s services and enable data-informed decisions, the program manager decided to create a data team to be responsible for the creation, maintenance, and reporting of *NewProgram*'s data.

So, before the launch of *NewProgram*, the data team met to start identifying the need for data creation. As a result of the meeting, the team drafted the following flowchart to define what data needs to be collected to answer questions about *NewProgram*'s operations and success.



Note that this is only part of the Identify Need step. Many other needs can be identified, and crucial questions must be answered for a strong groundwork. However, these are the first steps. If *NewProgram*'s team had skipped the Identify Need step and started to collect data unintentionally, it might have missed important information and been unable to perform certain analyses.

Plan

In the Identify Need step, you developed an understanding of why you should gather data and envisioned what that creation would look like. Based on this vision, you can start making practical decisions about your data creation process, such as defining which resources you will be using (e.g., people, tools, software, and budget) or addressing data privacy concerns.

The first consideration you need to have in mind when planning data creation is that the data needs to be usable. That means not only choosing the right data to create (i.e., the data that will answer your identified questions) but also creating it in a way that other people can use, update, and analyze.

Additionally, you should define the practices that will increase the efficiency of your creation process and reduce the chance for errors and unexpected barriers, such as not being able to use a specific software because of a license issue.

Here are some guiding questions that can help you make the right practical decisions:

Who will be involved in the data project?
• Who will be part of the project team? • Which additional people, voices, and stakeholders (e.g., community members, boards and commissions, staff from other agencies, etc.) will be involved? • What responsibilities will those involved have? What is the time commitment for each of them? When will they get involved? • How will the relationship with stakeholders be established?
What tools or software will you need to work with the data?
• What do you need to do to have access to it? • Who will be using them? Will they need to be trained? • Can the data be easily transferred from one system to another? How will this affect your or future projects?
How can you perform your project within your budget?
• What is your budget? What limitations does it impose on your project? (E.g., selected software is not within budget.) • Is there a way to overcome these limitations? (E.g., find another software or request more funding.)
Does the data have any special privacy requirements? (e.g., personal data, protected health data)
• What can you do to ensure data privacy? Who will be allowed to handle the data? • Will you or your team need to sign agreements or contracts prior to handling protected data? • Will you or your team need to be trained on handling protected data?
Does the data have other mandatory requirements?
• Which internal mandatory requirements do you need to follow? For example, will your collected data need to follow the City's Records Retention & Disposition Schedule? • Which mandatory requirements from external entities (e.g., federal or state government) do you need to follow? • Which resources should you seek to ensure compliance with identified requirements?

Equity Considerations: Plan

The Plan step is the structural base for all the following steps in the Create phase. For that reason, you need to make sure that your practical decisions consider equity to reduce the risk for bias and unfairness in the data creation process. Some actions that you should take in this step are listed below:

Ensure your team and other people involved in the project are diverse and inclusive. Reach out to people from diverse backgrounds to make up your team, contribute as consultants, and generally get involved as stakeholders in the creation process. By ensuring the inclusion of people from diverse ethnicities, races, cultures, genders, ages, education and work experiences, job titles, and City agencies, you will reduce the risk for implicit bias and unfairness.

Remember to include non-protected data in your plan for addressing data privacy. Data not protected by statute might still contain sensitive information. Identifying who will be allowed to handle the data and which agreements, contracts, or training will be required prior data handling are crucial actions for dealing with protected data. However, you should make sure your team also understands and is prepared to handle non-protected, but still sensitive, data that still could present a risk for a person's privacy, especially those who are part of a minority group.

If using existing data, plan how you will address any potential limitations or impact on equity. You can start by trying to find why people collected the data you want to use and how they collected it to assess if it was built on equitable principles. Even if you cannot find this information or found concerns, you can still use the data if you take the time to understand its shortcomings and plan how you can compensate its limitations.

If working with existing data, how will you obtain it?
• Is it available to the public, or do you need to request it? • If you need to request it, will you need to build a new relationship with the data owners? Will formal agreements be needed? • How long will it take for you to have access to it?
How will you ensure your plan is comprehensive and specific?
• What process should you take to compare your drafted plan against the needs for data creation identified in the previous step? • Which specific details need to be defined for the particularities of your project?

Sidebar 3: *Project Charter*

A **project charter** is a formal, typically short document that describes your project in its entirety – including what the objectives are, how it will be carried out, and who the stakeholders are (Wrike).

You can use project charters to document the practical decisions made in the Plan step, including but not limited to the following:

Project Vision	**Project Team**	**Problem Statement**	**Budget**
Roles & Responsibilities	**Scope of Work**	**Deliverables**	**Resources**
Objectives	**Stakeholders**	**Timeline**	**Privacy Concerns**

Additionally, project charters are useful for keeping track of what is in the scope of your project and effectively communicating all of your plans and expectations to stakeholders. In fact, project charters, or similar documents, are strongly recommended when working with multiple stakeholders to facilitate communication and ensure all parties are on the same page. Still, they are also valuable even when working with small internal projects.

A project charter template is available in Appendix C: Project Charter Template. However, your agency may have a recommended internal template, or you may use any other preferred source.

Another important point to consider is where and how you will store your data. You may be using a dedicated software program that handles data organization and storage for you, such as Accela or Cityworks. In fact, the use of existing and well-known software or enterprise systems is preferred. This will reduce your budgetary needs, make it easier for other data users and practitioners (i.e., people will not have to learn a new system), and allow for more consistent processes.

A Data Journey: Plan

After identifying the need for data creation, *NewProgram*'s data team started to draft a project charter to document the practical decisions about the entire data creation process. It included naming all project team members, defining the problem statement and background, addressing privacy concerns, identifying project resources, and specifying the members' roles and responsibilities.

The following image shows the first draft of the project charter:

Project Name: *NewProgram*'s Data Collection

1.0: Project Team	
Project Lead	Dusty Masterson, Public Health Department
Project Team	Jamie Duffy, Public Health Department Kris Napoletani, Public Health Department Hira Anthonsen, Fire Department
Project Sponsor(s)	Joss Cheung, Public Health Department

2.0 Problem Statement & Background

This project has the goal of collecting data about *NewProgram*'s operation.

3.0 Privacy Protocol

Because of the sensitive nature of NewProgram's data, the team will ensure that the data is protected by storing it in a network drive location only accessible by the project lead and team.

4.0 Scope of Work

5.0 Project Resources

Data

- No existing data is available. *NewProgram*'s team will collect all data through the course of its work (e.g., filling out forms as a step when responding to incidents)

Software and Tools

- *NewProgram*'s team members will be trained to use *ToolName* to collect data
- *NewProgram's* team members will use *SoftwareName* to store data

6.0 Final Product

7.0 Roles & Responsibilities

Role	Description	Person	Estimated Time
Team Leader	Supervisory support of data collection Regular communication with stakeholders	Dusty Masterson	1 hour per day + 1 additional hour per month (monthly meetings with sponsors and other stakeholders)
Team Members	Collection of data	Jamie Duffy Kris Napoletani Hira Anthonsen	3 hours per day
Sponsors	Support NewProgram's operations, including data collection Provide guidance	Joss Cheung	1 hour per month (monthly meetings with team leader and other stakeholders)

8.0 Timeline

Deliverable	People	Due Date

Note that the project charter still needs to address other crucial Plan questions, such as specifying the scope of work, identifying stakeholders, and defining a project timeline.

However, if you are managing the data yourself, some best practices to follow are:

- **Be consistent**: Make a consistent storage plan that you will keep throughout your data creation process. For example, if you decide to keep your files in a specific folder in a common drive, try not to move them around inside the drive or completely change storage location, unless you have good reasons. Also, aim for consistency in every decision you make (e.g., settings, name conventions, layout, formats, systems, and other plans).
- **Make your data accessible while following privacy requirements**: Define settings and naming conventions that will ensure both the security and accessibility of your data. That is, everyone who is allowed to access your data should be able to take an easy, intuitive path to it.
- **Organize your data in a tabular format**: Plan to store your data in the form of a table with rows and columns. Tables are a neat and convenient way to present a large body of information that includes repeating data elements (Techwalla). Using a dedicated data program is preferred; for instance, use Excel spreadsheets instead of Word tables.
- **Strive for file formats that will be useful for current and future projects**: Whenever possible, plan to store your data in a machine-readable format. That is, a format that can be automatically processed by a computer, such as CSV, JSON, XML, etc. (Open Data Handbook)
- **Plan your backups**: Think about how you will protect your data from accidental loss by defining how and when to perform regular backups. Check with your agency to learn about its standard or preferred backup methods.
- **Maintain a single source of truth**: Define how you will ensure that your project will produce only one source that provides the right data points to be used for decision-making. In other words, plan how you will avoid having different sources that may hold conflicting information. This is important because it reduces the amount of time spent on identifying which record is the most updated or correct, and it ensures that everyone making decisions will use the same data.
- **Define update procedures**: Old versions of data can provide valuable information but can also be cumbersome to maintain. When deciding whether to keep or overwrite data values, think about what is practical now while considering the current and future benefits of retaining past data. Additionally, some datasets may follow the City's Records Retention & Disposition Schedule as established by the APM 3-6: Records Management Program.

After finishing planning the technical aspects of your data creation work, you should establish core concepts of quality assurance that will be applicable to the lifecycle of data management, as demonstrated in the next step.

This page left intentionally blank

Quality Assurance

Before learning how to establish good Quality Assurance practices, it is important that you know how to measure the quality of a data or dataset. **Data quality** consists of five measurable elements:

Accuracy	Accurate data and datasets truly reflect what actually occurred, and the math behind them is correct and logical.
Completeness	Complete data and datasets are appropriately filled out, and City Agencies have records on everything appropriate.
Consistency	Consistent data and datasets stored in different locations or processed by different people should have identical processes and structures without conflicting with each other.
Validity	Valid data and datasets truly measure what you intend to measure (e.g., number of complaints is not a valid measurement of resident satisfaction because not everyone knows how or has the time to file complaints).
Verifiability	Verified data and datasets are aligned with an existing, verifiable source.

Your data is considered quality data when it measures high in all five elements. This is important because quality data is the foundation for better decisions, more accurate stories that demonstrate the collective impact of City services, and better auditing.

So, how can you ensure that your data will be accurate, complete, consistent, valid, and verified? The answer starts with Quality Assurance. It forms the data quality management process together with the Quality Control and Data Cleaning steps, which come later.

Quality Assurance (QA) is the process of identifying and applying **proactive** methods to prevent errors and ensure that your project meets data quality standards.

In this step, even before the data is gathered, you should start by creating an outline for how to review it. This will help you think systematically about the kinds of errors, conflicts, and other data problems you are likely to encounter in a given data set (DataOne). Then, you can establish proactive methods to minimize the occurrence of these problems.

Additionally, when developing your QA practices, follow the City's **data standards** outlined in Appendix B: City of Madison Data Standards whenever possible. The goal of data standards is to make data consumption easier for all users and reduce the chances of inconsistent and inaccurate data. They include what information needs to be entered and in what consistent format (e.g., address format, numbers, percentage, etc.).

Examples of Quality Assurance Practices

- Using automated processes to run tasks that are performed on a schedule or multiple times, such as SQL queries, Excel macros, FME routines, Visual Basic or Python scripts.
- Setting up drop-down lists to control input values, such as only allowing users to select from a list of City agencies instead of typing the agency name in a free text field.
- Creating rules to reject the entry of invalid values, such as a day of birth that is set in the future.
- Limiting inputs to certain data types, such as only accepting numbers for a field representing recycling tonnage or dates for a field representing program start date.
- Creating duplicate detection rules, such as flagging when a similar entry already exists in the database and forbidding the creation of the duplicate or suggesting merging the entries.

Equity Considerations: Quality Assurance

In this step, you will define what is good and bad data to create quality assurance practices. It is natural that you will use your personal and work experience to create your definitions. This approach is helpful, but it also gives a chance for your biases to enter the data. See below how you can avoid that.

Be careful of discriminatory quality assurance practices. In this step, you learned that setting up domains and limiting inputs are ways to ensure data quality. But, be careful not to limit inputs only based on your world view and experiences. For instance:

- Creating a dropdown list for races and ethnicities without adding an option for "others" can exclude groups and individuals who do not identify with any of the options in the list.
- Limiting the number of letters in a last name field, such as "at least three letters" or "no more than fifteen letters," can exclude individuals who do not have common U.S. last names. For example, many immigrants have two or more names in their last names while others have two-letter last names.
- Making an address field required without providing options for persons experiencing homelessness.
- Assuming that all children have two opposite-sex parents, such as asking for information about "mother" and "father" instead of more neutral terms like "parent/guardian 1" and "parent/guardian 2 (if applicable)".

Talk to a diversity of people before setting up your quality assurance practices. Reach out to people from diverse backgrounds to help you define your quality measures. Make sure you include people from diverse ethnicities, races, cultures, genders, ages, education and work experiences, job titles, and City agencies.

Check the data of other similar projects. Find similar projects to get insights about what kinds of data you might expect. This can help you expand your views and build more inclusive quality assurance practices.

Some questions that can help you define which Quality Assurance practices are most appropriate for your project are:

- How will you make sure that your data will tell the right story?
- How will you prevent impossible values of data from being recorded? How will you double-check unlikely ones?
- What can you do to ensure that your calculations will be logical and correct?
- What will you do about missing data?
- Are there City standards for the type(s) of data that will be produced? If not, how can you ensure your data will not conflict with data processed by different teams or agencies?
- Are there state or federal requirements you must follow? What about other entities' standards, such as industry leaders?
- How will you ensure that your data is measuring what you intend to measure?
- What verifiable sources can you use to confirm the accuracy of your data?
- Do you have available data to test your QA practices against, such as data from a previous year or comparable project?
- Which practices do your tools or software allow? How will you deal with tools or software limitations?
- Who will be responsible for monitoring data quality and verifying assurance practices? Who will take action if quality issues are found?
- Which additional standards and guidelines should you develop for your specific project? For example, consider your project collects data on dead animal locations. Your form asks residents to type in a text field either the street address or the intersection where the dead animal was found. Besides following the directions for street addresses presented in the City's data standards, you will have to define how to store intersection information in a clear and consistent way. Other situations may involve dealing with ambiguous survey answers or narratives. If you predict that ambiguity may occur, think about what you can do to minimize its occurrence (e.g., providing examples to answers) and address it after collection (e.g., reaching out to data collectors to clarify the ambiguity).

Another essential part of Quality Assurance is to record the critical decision points of your work in order to establish a replicable process by others now and in the future. So, in this step, you also need to identify how you and others will document your decisions and actions, which might include comprehensive documents, such as detailed field notes, or more concise and structured annotation, such as metadata. See sidebars 4 and 5 for more information on data documentation and metadata.

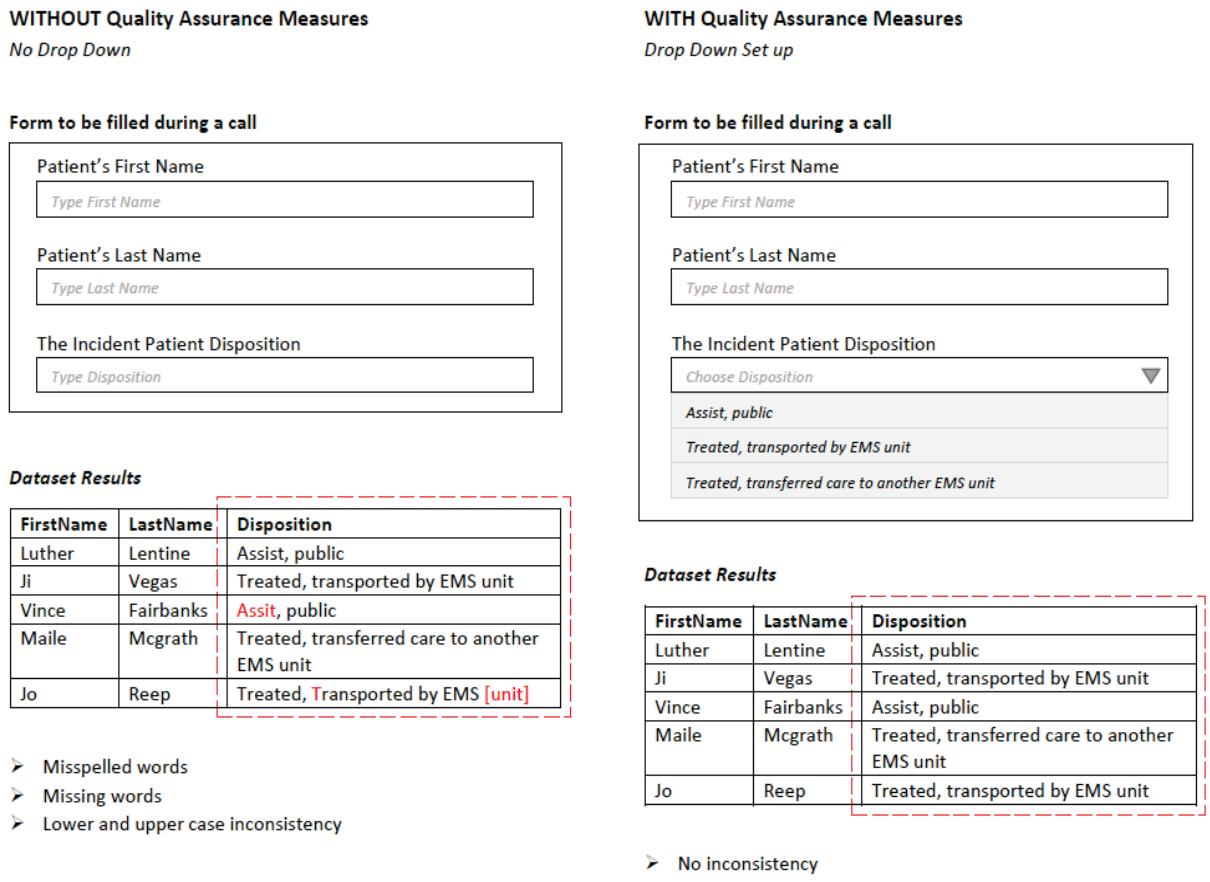
In summary, Quality Assurance is all about **proactively** ensuring the creation, maintenance, and reporting of quality data. It is not a standalone activity done only in the Create phase. You will perform it throughout the entire data management process. However, while there are differences in how QA is approached at different stages, you will continue to put into practice the core concepts of error prevention and documentation.

A Data Journey: Quality Assurance

NewProgram's data team decided to use drop-down lists to control the values of certain inputs in the program's forms as one of its Quality Assurance practices. For example, one of the fields chosen for drop-down setup refers to the patient disposition in the *NewProgram*'s incident form.

As *NewProgram*'s frontline workers manually fill out the form at the moment of the response, there is a lot of room for inconsistencies and errors. But because there are only three possible dispositions in the program, *NewProgram*'s data team can create a drop-down list of dispositions, reducing the occurrence of data quality issues.

The following image compares NewProgram's possible future dataset results with and without the QA measure mentioned above:



Note that using a drop-down list was an appropriate measure to assure consistency in the disposition field. However, Quality Assurance methods will vary from case to case.

Sidebar 4: Data Documentation

Data documentation describes and contextualizes your data to ensure that it will be understood and interpreted by any user, including your future self. Besides improving interpretation, it can be especially helpful for new staff training, ensuring effective quality procedures, and reminding you of previous steps in the project. Some examples of documentation are readme files, screenshots displaying settings, and metadata.

Your documentation does not need to be extensive or drawn out, but it should include the following information about your data: *(1) How it was created, (2) Where it comes from, (3) How often it is updated, (4) Description of its context, structure, and contents, (5) Any assumptions you made, (6) Any limitations you identified, (7) Any manipulations/calculations that have been done to it, and (8) All the data quality management steps taken and applied to it.*

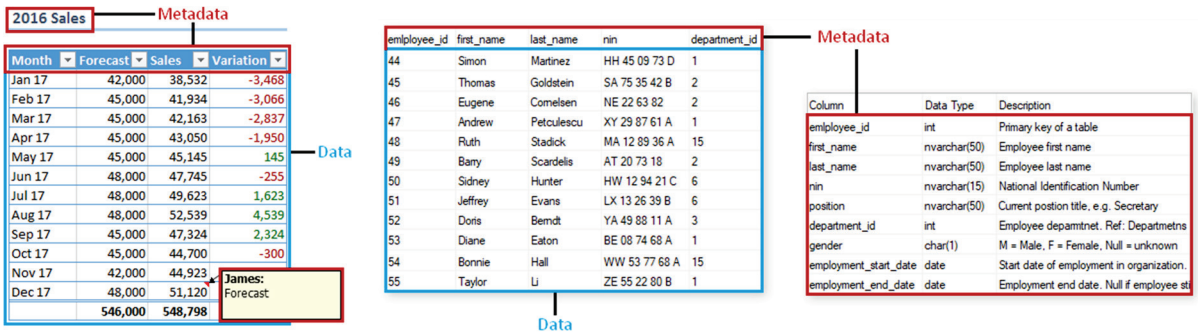
At this moment, the City of Madison does not have a cross-city standard for the content of these documents nor how they are saved. However, you should name and store them in such a way that they are easy to locate and link to the data they concern. Check with your agency for any department-specific standards.

Source: Axiom Data Science and the University of Arizona.

Sidebar 5: Metadata

Metadata is a form of data documentation that summarizes information about data in a concise, structured way. There are many distinct types of metadata, such as:

- **Descriptive metadata**: Descriptive information such as title, abstract, author, and keywords.
- **Structural metadata**: Information about how the data is organized. For example, a table of contents, the number of file versions, and the relationships and other characteristics of digital materials.
- **Administrative metadata:** Information to help manage a resource, like resource type, permissions, and when and how it was created.
- **Reference metadata**: Information about the contents and quality of statistical data.
- **Statistical metadata**: Statistical information about the data, such as processes and calculations that were used in the production of statistical data. Sometimes also referred to as Process Data.
- **Legal metadata**: Provides information about the creator, copyright holder, and public licensing, if provided.



Source: Dataedo, Towards Data Science, and Wikipedia.

This page left intentionally blank

Collect

With proper Quality Assurance mechanisms in place, you are ready to begin gathering your data. As you have gone through the past three steps, you likely have a good idea if you will need to collect new data or get it from an existing source (internal or external). Existing data simply needs to be accessed and entered, which means you can skip this step and move on to the Enter step. However, if you need to collect new data, this step will guide you on how to collect quality data to support decision-making.

First, you should start by identifying which data collection method(s) will be the most appropriate for your data needs. The following table shows the most applicable methods to City agencies.

Method	Overview
Interviews	• Useful for when you need to gain an in-depth understanding of perceptions and opinions on a topic, or for working in areas that are hard to objectively measure. • Your questions should encourage open-ended responses. • Your data will be mainly qualitative (e.g., narratives). • *City examples:* Participants' perceptions of increase in leadership abilities (*CDD*).
Surveys and Questionnaires	• Useful for when you need to understand the general characteristics or opinions of a group. • Your questions should include a set list of responses, such as a Likert-type scale (e.g., "disagree, neither agree nor disagree, agree"). They may also include some short, open-ended responses, but too many will be hard to interpret. • You will be able to analyze the responses with quantitative methods by assigning numerical values. • *City examples:* Customer satisfaction (*Streets*), community literacy levels (*Library*).
Observations	• Useful for when you need to understand something in its natural setting. •You can perform human observations (e.g., visual count of behavior) or machine observations (e.g., data collected through sensor devices). • Your data can be qualitative (e.g., narratives) or quantitative (e.g., frequency counts). • *City examples:* Number of people using spaces (*Library*), counts of patrons to beaches (*Parks*), speed of vehicles (*Traffic Engineering*).
Measurements	• Useful for when you need to measure the physical characteristics of something, such as an object's dimensions or the number of items in a group. • People can collect measurements by using measurement tools, such as scales and rulers. Or they can set machines up to automatically measure and record an object's measurements (e.g., weight scale, GPS transponder). • Your data will be quantitative (e.g., object's weight). • *City examples:* Truck weight (*Streets*), linear feet of storm pipe cleaned (*Engineering*).
Documents and Records	• Useful for when you need to gather information from documents and records, such as invoices, meeting minutes, attendance logs, and reports. • Can be an inexpensive collection method but may be an incomplete data source. • Your data can be qualitative (e.g., narratives) or quantitative (e.g., frequency counts). • *City examples:* Number of children signed up for summer reading (*Library*), number of polls that are open late (*Clerk*).

Source: CYFAR | University of Minnesota and Scribbr.

Equity Considerations: Collect

In this step, you need to make sure that your collection methods are not excluding and hurting groups or individuals. Here are some ways:

Design interviews, questionnaires, and surveys that are accessible, inviting, inclusive, and unbiased. For example, you should:

- Use inclusive and plain language and, whenever appropriate, explain concepts that your audience may not be familiar with;
- Choose color palettes suitable for color blindness;
- Phrase your questions in an unbiased way to avoid influencing people's responses;
- Cover topics and offer response options that will give respondents a way to fully reflect their experiences;
- Translate materials to other languages, when appropriate;
- Find alternative ways for your outreach, focusing on hard to reach populations, when possible.

Check the City's Content, Accessibility & Plain Language Tip Sheet and Gender-Inclusive Language Style Guide for more directions.

Recognize that "there is no such thing as raw number". We construct numbers by making decisions about how to separate things into groups (Deborah Stone). For example, consider that you are using the observation method to collect data on park usage. The first decision you make is to define three categories of usage: walking, jogging/running, and biking. Note that you decided that jogging and running are similar enough to be together. Other collectors could disagree with you, which would result in different numbers. In addition, you will likely have to make other decisions during the collection. For instance, where would you put someone who is walking, but is wearing athletic clothing and seems tired and sweaty? What about someone in a wheelchair? In situations like this, you will have to use your best judgment, which could be adding your biases to the data. For example: you could think someone walking in expensive workout gear must be taking a break from running, while someone running in more ordinary clothing must not really be there to exercise - maybe they are just trying to get quickly from one place to another – so their running does not really count. On an individual level, this may not seem like a big deal, but in aggregate, this bias may lead you to conclude that wealthier communities also run more, creating data that supports offering wealthier areas more running opportunities and structures than other areas.

Think about the data your method does not collect. For example, consider that your data collection relies on people using an online form to report a problem. At a first glance, you can think that this method will capture all problems in the city. But that is not entirely true. It could be excluding problems happening to people who do not have internet access, struggle with technology, or simply do not know the online form exists.

Be careful when collecting data from a sample. A sample is a portion of a larger group. When it is not possible to survey all members of the whole group, you can use samples to generalize information. But be careful. Often, the people included in surveys are the easiest to reach - who tend to be older, whiter, and wealthier than the general population. As a result, individuals who are part of excluded groups will be misrepresented by the data. Tools from the field of statistics can help mitigate – but not undo – these effects, so it is best to get the most representative data possible.

Regardless of which data collection method you choose, you should research and follow industry best practices for implementing it, particularly those with high human involvement. (E.g., what are the best practices for creating good surveys?)

You will also have to think about which approach you will take: manual or automated collection. Manual data collection is performed when people manually write, process, or download records (e.g., writing down interview notes, tallying users in a space, or collecting invoice papers). On the other hand, automated data collection is performed when people rely on tools that automatically capture and store data (e.g., online survey forms, sensor devices, scripts for online search).

Another aspect of identifying the right method is understanding what resources you will need to perform it, and if you will have access to these resources. Some questions that may help you with this task are:

- Will you need specific equipment or tools for the collection? Do you have access to them? If not, do you have a clear pathway to get access?
- What other resources and permissions will you need in order to collect the data?
- How many people need to be involved in the collection process? Who are they? Will they need training?
- Can you perform the method within your budget? If not, can you, or should you, get more funding? What would be the second best option?

Not only should you consider how you will collect the data, but you should also identify what categories of data you will collect. That is because the type of data collected influences the type of analyses and presentations that can be conducted later on. The following table outlines the main characteristics and uses of three categories of data.

	Discrete Data	**Continuous data**	**Nominal data**
Used when you are…	**Counting** something	**Measuring** something	**Classifying** something
Clues to look out for are…	• If you are counting whole things (i.e., things that cannot be subdivided, measured as part of a unit). *Example*: the number of cars in a parking lot.	• If you are measuring on a continuum or standard scale that can be infinitely divided. *Example*: the temperature of a lake.	• If you are categorizing things into labels that do not have any numerical value or order. *Example*: the gender of employees in an Agency.
Some common statistics are…	• Percentage (Proportion) • Ratio • Rates	• Average (Mean) • Percentage Change • Standard Deviation	• Mode • Percentage (Proportion) • Ratio
Some useful graphs are…	• Bar Graphs • Line Graphs • Heat Map • Pie Charts	• Bar Graphs • Line Graphs • Histogram • Gauge Charts	• Pie Charts • Pareto Chart • Stacked Bar Graphs • Tree Map

Source: Chi Squared Innovations and The Drum.

A Data Journey: Collect

In order to measure the quality of *NewProgram*'s services, the team has identified a need to collect data on patient satisfaction after an incident response. The first draft of the incident form did not provide a specific field for entering patient satisfaction, and frontline workers could only write patient comments about satisfaction in the Patient Notes field. However, after evaluating the form against data collection considerations, the data team noticed that collecting satisfaction data through such a field would not be efficient nor consistent. Thus, the team decided to include a field for a short survey question at the end of the form.

The image below compares the collection of patient satisfaction data via the first draft (i.e., Patient Notes field) and the new draft (i.e., specific field added for patient satisfaction):

WITHOUT collection considerations
Patient Service Complaints typed on Notes

Patient Notes (Included in the call log form)

Patient initially refused to inform his address and other important information. He was suspicious and disoriented. Patient informed dissatisfaction with the way he was being treated. After 20 minutes, our team was able to control the situation and transported the patient to his preferred health care institution.

- Insufficient information: hard to identify when patient is satisfied or indifferent with *NewProgram*'s services (most of the time only dissatisfaction will be noted).
- Filed mixes information about patient satisfaction with other information: resources will be wasted trying to review all notes for any indication of patient satisfaction and dissatisfaction.
- Complicates and delays data entry process.

WITH collection considerations
Survey for Patient Service Satisfaction

Survey Presented to the Patient after Treatment

Please, rate our services:

☐ Good ☐ Average ☐ Poor

- Easier and clear way to collect and classify patients' levels of satisfaction with *NewProgram*'s services.
- Less wasted resources.
- Makes data entry process easier and faster.

In addition to identifying which category your data belongs to, you should think about its **data type**. A data point's data type tells a computer system how to interpret its value. For example, formatting an Excel cell as a date allows to you perform date-specific operations, like formatting its display and finding the number of days between two dates, and formatting digits as numbers instead of text allows you to easily perform mathematical calculations. Thinking about data types before collection will help you collect data in a way that will require less manipulation during the Enter step.

Moreover, it is important to collect data at consistent intervals to improve its reliability. For example, Job Family Availability Data (JFAD) collected, entered, and reported annually can vary based on the time of year due to seasonal employees. To address this concern, the Division of Civil Rights (DCR) worked with the Racial Equity and Social Justice Initiative (RESJI) Data Team and Data Governance Team members to understand this data and create better reporting on a monthly basis that separates out permanent and seasonal employees to capture seasonal variation.

It is also crucial to ensure that every person working in this step is familiar with data privacy categories and their limitations. These are discussed more fully in the Data Privacy section, but here is a summarized version:

> **Public** - This data can be publicly disseminated without any concerns.
> **Protected** - This data is protected by law or regulation and can only be shared or accessed by a limited group or through a limiting procedure; if cleaned to remove certain information, or aggregated, this data could potentially be shared.
> **Sensitive** - This data is not regulated like protected data, but in its raw form, this data poses security concerns and could potentially target individuals or pose other concerns; if cleaned to remove certain information, or aggregated, this data could possibly be shared.

Lastly, to avoid accidental loss of data, you should back it up at regular frequencies, including when you complete your data collection activity and any time you make edits.

This page left intentionally blank

Enter

Data entry is defined as inputting data or information into a computer using devices such as a keyboard, scanner, disk, or voice technology (Computer Hope) by using a software program, such as Excel, Accela, or MUNIS. Data is important to our operational effectiveness, decision-making, and customer service delivery. It is most likely that your agencies regularly require some type of data entry, like financial figures, email addresses, operational records, client names, or meeting minute transcriptions.

The ultimate goal of data entry is to create an organized and useable set of data in a specified format. High-quality data entry is the foundation of accurate decision-making, a critical part of the success of our entire data management process, and an important goal for us to achieve. Your entire entry process and its outcome should be focused on the achievement of this goal. For that, you must ensure that the data is entered properly to avoid errors and inconsistencies and minimize the time needed to clean it later in the process.

Here are some specific steps (and cautions) for successful data entry:

Ensure data standards are in place

As mentioned in the Quality Assurance step, your entered data should follow the City's data standards (outlined in Appendix B: City of Madison Data Standards) to ensure consistency between and within datasets.
However, in instances where there are no specific standards for your data or guidelines are not operational, you could verify if other agency-level standards fit your needs or create additional guidelines for your project. **But remember to document your decision as part of your QA practices.**

Assign descriptive names to files and columns

A good practice in table-based data entry is to create descriptive names for files and columns. You may include details such as source, date, version, project, etc.
Additionally, you should not include spaces or special characters within them because it may create problems when the data file is used for subsequent analysis.

Know how to address missing data

Most of the time, datasets end up having missing data. This can lead to significant problems if they are not identified and located during data entry.
There are several ways for you to minimize the impact of missing data that will be used in future analyses. Start by checking the City's data standards.

Keep a log

Keeping a log is vital when you are carrying out data entry. A log contains the record of errors and difficulties that you encountered in the entry process. It provides a systematic account of the process efficiency and can be useful in fine-tuning the data entry process and project management.
For example, you may record the number of fields from which information is missing; fields in which wrong and inaccurate data has been entered; fields that need clarification; when an error was noted; and when an action was taken. Your log should be clear and saved in a standard location.

Equity Considerations: Enter

Sometimes this step simply involves copying data without changing it, such as from a paper form to a database. Sometimes, however, it also involves interpretation. In this case it is crucial not to let your biases affect your work, especially around subjective topics.

Do not let your biases enter the data. First, make sure you know your biases (see Equity Considerations: Identify Need). Then, identify if you are passing them into your data. For example, in the Collect phase, you may have recorded interviews, and in this phase you are interpreting them to record the subjects' mood. If you have a racial bias towards Black women, who have sometimes been stereotyped as aggressive (Angry Black woman stereotype), you may judge two similar answers as angry for a Black respondent and neutral for a White respondent.

Verify your assumptions. Sometimes your data entry work may involve some level of interpretation. For example, you may enter data by transferring it from a paper form to a spreadsheet and have to decipher someone else's handwriting, or, as in the example above, you are making subjective inferences like moods. In such a case, you may face uncertainty and rely on assumptions based on your experience with the data and your understanding of the real world the data reflects. That is an opportunity for your biases to enter your data work. A way to avoid that is to verify your assumptions, such as reaching out to the person who filled out the form to confirm the entry. In some cases, you may not be able to verify your assumptions with the source, so make sure you document what the assumptions are and how you developed them.

Automate

When possible and reliable, use automation to carry out large volumes of data entry but remember that validation needs to be done in order to ensure accuracy (e.g., SurveyMonkey has automated question formats, collection, and analysis options that can be used). Copy and paste features can also help you avoid data errors. On a larger scale, IT can help move data between systems, locations, and formats

Train data entry staff

Data entry staff should be trained in recognizing and identifying common data entry errors, which can include mistakes within fields or between similar files. These errors cause difficulties when sorting and searching and make the dataset unusable. Some common mistakes are:

- Typos (e.g., Madisn-Madion-Masidon)
- Inconsistent formats for name, address, dates (e.g., Smith, John vs. John Smith)
- Different types of information in one column (e.g., city and state in one column)
- Wrong column order
- Missing data

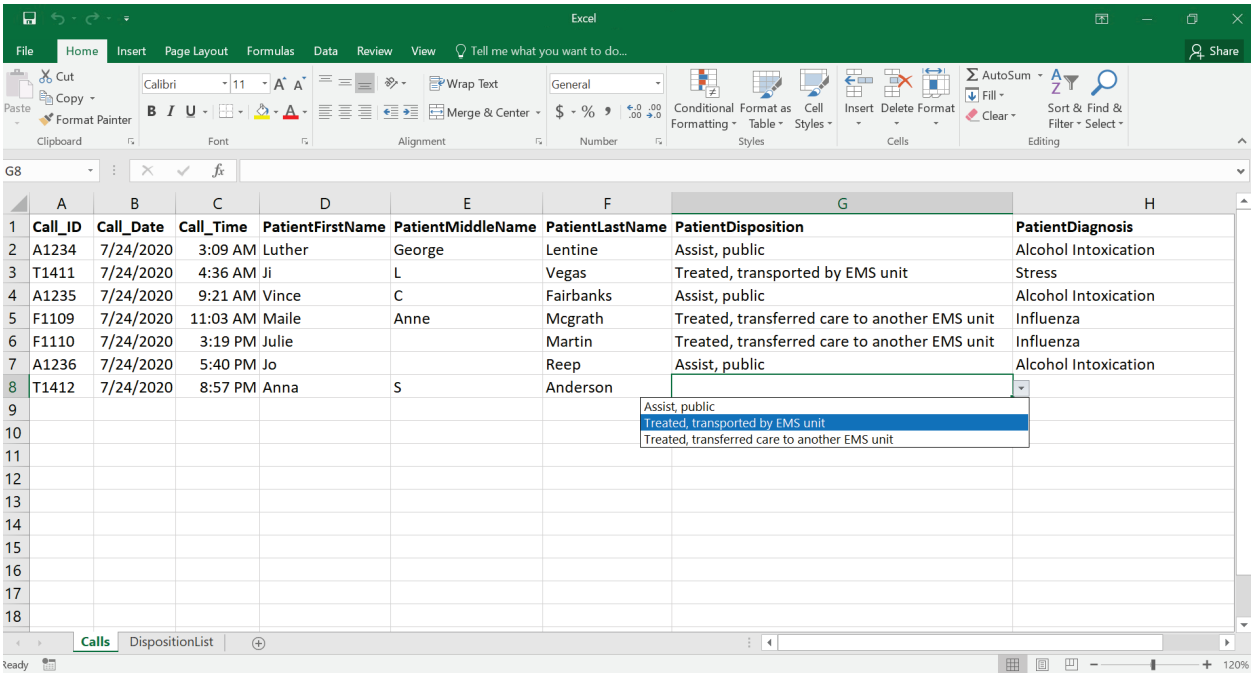
In summary, you can minimize the chance for error by establishing a thoughtful data entry process with strong Quality Assurance practices and ensuring that the entry staff follows these processes and practices. Additionally, you can reduce the number of different personnel entering data at different times or avoid data entries being made in different files to be combined later in a final dataset.

Even if you follow all the steps above, data entry mistakes and mishaps can still occur, which is why you will learn about quality control in the next section.

A Data Journey: Enter

NewProgram's team responsible for data entry uses Excel to enter the data collected by frontline workers. One of its datasets includes data on the date and time of the service, the name of the patient, the patient disposition, and the diagnosis at the moment of the incident.

The following image shows a *NewProgram*'s team member entering data for patient Anna Anderson (8th row).



	A	B	C	D	E	F	G	H
1	**Call_ID**	**Call_Date**	**Call_Time**	**PatientFirstName**	**PatientMiddleName**	**PatientLastName**	**PatientDisposition**	**PatientDiagnosis**
2	A1234	7/24/2020	3:09 AM	Luther	George	Lentine	Assist, public	Alcohol Intoxication
3	T1411	7/24/2020	4:36 AM	Ji	L	Vegas	Treated, transported by EMS unit	Stress
4	A1235	7/24/2020	9:21 AM	Vince	C	Fairbanks	Assist, public	Alcohol Intoxication
5	F1109	7/24/2020	11:03 AM	Maile	Anne	Mcgrath	Treated, transferred care to another EMS unit	Influenza
6	F1110	7/24/2020	3:19 PM	Julie		Martin	Treated, transferred care to another EMS unit	Influenza
7	A1236	7/24/2020	5:40 PM	Jo		Reep	Assist, public	Alcohol Intoxication
8	T1412	7/24/2020	8:57 PM	Anna	S	Anderson		

Note that the dataset follows the City of Madison Data Standards for date, time, and name of persons. Additionally, the PatientDisposition column provides a drop-down list to ensure that information is being entered correctly and consistently, in accordance with strong QA measures.

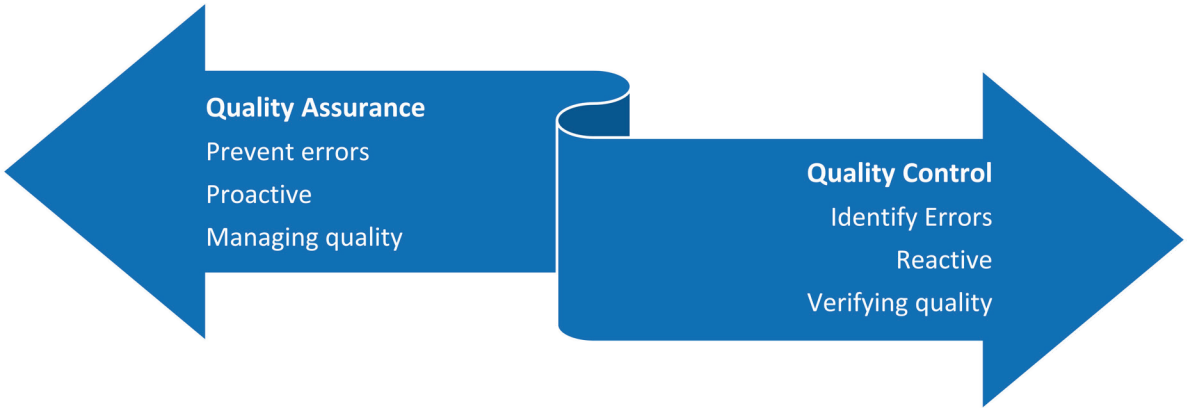
Quality Control

Quality Control (QC) is a system of routine and planned processes established to measure and control the quality of data. In this step, you will test your data against a known set of expectations, which are based on the five measures of quality: accuracy, completeness, consistency, validity, and verifiability. If quality issues are found, you will apply defined procedures to fix them. How often you should run your QC practices depends on the dataset, but they should be performed on a periodic routine basis, especially at the end of a major step and before you post or share your data.

Quality Assurance and Quality Control are often confused for each other. The following sidebar explains their differences:

Sidebar 6: *Quality Assurance and Quality Control*

Quality Assurance and **Quality Control** are two aspects of quality management and crucial steps to the creation, maintenance, and reporting of quality data. While both steps are systematic and focus on standardized procedures, they differ in the following ways:



Quality Assurance is first performed prior to data collection and entry. Its goal is to prevent errors, and it is a more proactive approach to maintaining data quality. It can be revisited and updated later in the management process when you routinely pick up the same errors during Quality Control.

Quality Control, on the other hand, is a reactive process that occurs after data entry and as an ongoing part of data maintenance. It is aimed at identifying and strategically fixing systematic errors and inconsistencies to ensure data quality.

What is the Quality Control routine process? QC techniques depend on the dataset but typically consist of screening for strange patterns or inconsistencies and diagnosing the data using analytical techniques to understand what may impact inaccurate analysis. As a City, we do not have a standardized QC routine process since most datasets will require individualized practices. Check with your agency if there is a recommended process in place.

Equity Considerations: Quality Control

In this step, you will use your experience to judge what is and is not "normal" or "acceptable" data. Although using your experience is certainly helpful, it also creates a chance for your biases to enter your work; or even, for you to apply your experience to situations where it does not actually apply. So, make sure you take other approaches to reduce these risks. See below for three ways to ensure unbiased Quality Control.

Make sure you have a "good reason" for your quality control practices. For example, consider you found that most entries for race/ethnicity are missing in your dataset. Then, you find that there is a model that predicts the probability of a person's race/ethnicity based on the person's name. On an individual level and for the Create phase, using this model would be a mistake because you would be relying on assumptions about the relationship between names and races/ethnicities. But in some cases where you would be using aggregate data for analysis (Report phase), you could use this model to include racial data. As you can note, unfortunately, there are no standards for what defines a "good reason"– it often varies from case to case, so you will have to use your best judgment, check with others, and always document your decisions.

Verify your assumptions before making any big changes. For example, consider your dataset has a column named "gender" with values 0 and 1. You cannot find the documentation about what which value represents. You could assume what they represent based on how it is usually assigned in your department or how you would assign them, but that could lead to a mistake. So, you should find ways to verify if your assumption is correct. Some ideas are to talk to people who may be able to clarify your doubt or compare the entries with other documents from the collect step. Also, make sure you document your assumptions and changes.

Talk to subject-matter experts and whoever worked on the previous data creation steps. Even if you do not spot an issue, you can reach out to these people. Because they are familiar with the data and its context, they could spot something wrong that may look right to you (according to your own assumptions).

As a data user, there are plenty of actions that you can include in your Quality Control routine:

	Within Record Check *Checking Samples*	Between Record Check *Summary Statistics*
Accurate	• Make sure every value passes your gut check (e.g., does it seem accurate to have "10" in a record for an employee's age?). • Compare your set to an original record where appropriate (e.g., the coding of an interview compared with a transcript of an interview).	• Look at ranges, averages, distinct categories (e.g., run queries on data to make sure nothing is out of range). • Use tools to detect duplicate records, outliers, or unexpected data.
Complete	• Look at each record to make sure it is complete (e.g., do you have all the required fields for employee A?)	• Look at total counts as well as counts by subgroup, such as by year or by geographic area (e.g., create histograms to check the distribution of values and look for gaps). • Check for changes over time (e.g., new data available).
Consistent	• Compare samples to each other (e.g., do similar underlying situations seem to have been handled in similar ways?).	• Verify if similar situations were captured the same way (e.g., do you have a set with 25 records for "MFD" and 1 for "Madison Fire"?)
Valid*	• Assess if your set usefully and realistically captures the situation you want to represent (e.g., check the criteria that were used to define what should be counted or not during collection). • When possible and appropriate, ask those who requested the data or are experts on the topic to check the validity of your data.	• Compare the collection tools used to your objectives (e.g., you can use RESJI's Tools to ensure your data has been collected in alignment with your equity objectives). • Check if comparable datasets are presenting similar stories. • Check for gaps in your summary statistics.
Verifiable	• Compare records to accepted sources (e.g., original records from earlier work, comparable records from other projects).	• Compare your summary statistics to accepted sources (e.g., summary statistics from an earlier stage of the same project or from comparable projects).

**Note that validity in the QC process can be hard to detect, and even harder to correct – if you realize you have collected invalid data, you will most likely have to return to the Collect step to collect valid data. Ideally, you should think about validity during the QA step to ensure you will collect valid data. However, you can make some attempts to detect it during QC by closely examining your data, both individual samples and as a whole, by thinking about the situations you hope to capture and seeing if it does, and by looking for unexpected gaps.*

No matter the nature of the issue spotted, if your Quality Control routinely picks up the same errors, you should add preventive routines for that to your Quality Assurance practices, whenever possible. Additionally, you should make sure you document the steps and decisions taken, including any modifications to the Quality Assurance practices.

A Data Journey: Quality Control

Jamie, a member of *NewProgram*'s data team, was assigned to regularly verify the quality of the dataset below by performing set Quality Control routine practices, which include looking for duplicated entries, missing values, and potential misspells.

Call_ID	Call_Date	Call_Time	FirstName	LastName	
A1234	7/24/2020	3:09 AM	Luther	Lentine	
T1411	7/24/2020	4:36 AM	Ji	Vegas	1
A1235	7/24/2020	9:21 AM	Vince	Fairbanks	
A1235	7/24/2020	9:21 AM	Vince	Fairbanks	
A1235	7/24/2020	9:21 AM	Vince	Fairbanks	
F1109	7/24/2020	11:03 AM	Maile	Mcgrath	2
F1110	7/24/2020	3:19 PM	Julie		
A1236	7/24/2020	5:40 PM	Jo	Reep	3
T1412	7/24/2020	8:57 PM	Anna	Anderosn	

After applying methods to highlight duplicates in Excel, Jamie found three duplicated rows (1). Additionally, they filtered each column to look for missing values, finding a blank last name cell in the call F1110 (2). Lastly, they compared the first name and last name columns with the equivalent columns in another dataset, which returned an abnormality in the spelling of the patient's last name in the call T1412 (3).

Then, Jamie investigated all the potential issues mentioned above to ensure they were true errors:

Issue 1 – duplicated entries: Jamie found that there was a system malfunction at the time of entry, generating three identical rows. In order to clean this, they deleted two of the duplicated rows.

Issue 2 – missing last name: After following up with the responder who filled in Julie's information, Jamie found that the responder could not get Julie's last name at the moment of the response. In order to clean this, Jane followed up with Julie to update the record. Additionally, they identified that this error could have been prevented by adding a new measure to *NewProgram*'s QA practices, such as making last name a required field.

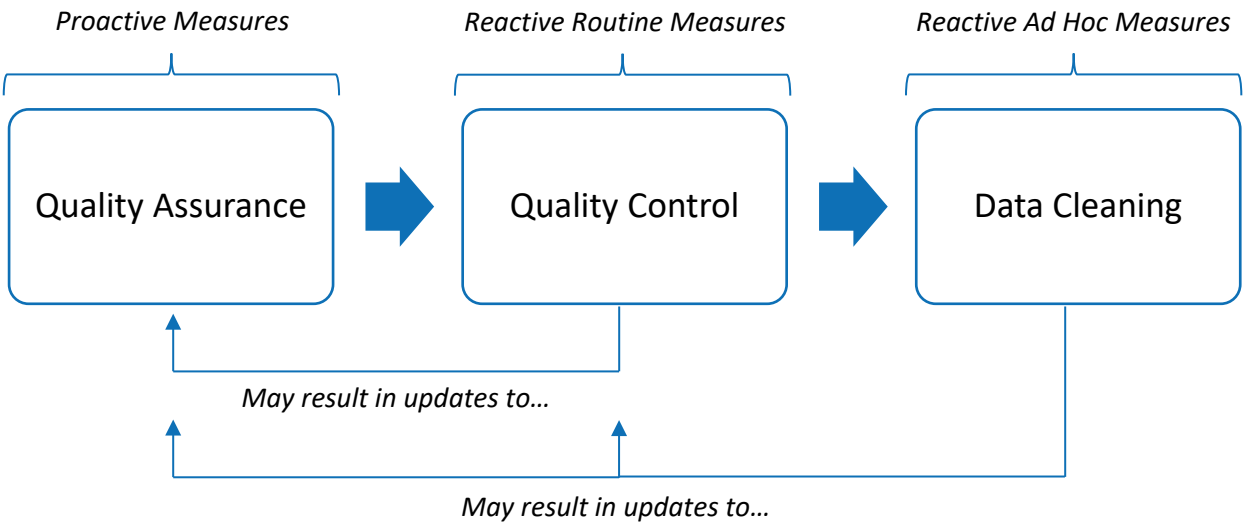
Issue 3 – potentially misspelled last name: In order to confirm Anna's last name, Jamie checked this dataset against another one provided by *LocalHospital*, *NewProgram*'s partner. They found that Anna's last name is Anderson, correcting it on the dataset.

If *NewProgram*'s team had skipped the Quality Control step, there could have been consequences at the individual and aggregated levels. For instance, at the individual level, the team could have used Anna's misspelled last name in official documents, resulting in legal issues or rework. And, at the aggregated level, the team could wrongly conclude that there were nine calls on July 24, 2020. Moreover, they could think that most calls are happening around 9 am and design a program/policy based on inaccurate information.

Data Cleaning

Data Cleaning is another important step in the data quality management process. In this step, you will continue to improve the quality of existing data by detecting and correcting issues, but in a less systematized way than your QC routines. In other words, you will perform the same actions as presented in the Quality Control section, but only when you identify an occasional need for it.

Because of this ad hoc nature, the Data Cleaning step is intended for one-off issues, that is, errors and inconsistencies that are infrequent and isolated to the point of slipping through QA and QC processes. However, it is important to note that the strength of your QA measures and QC routines will determine the amount and type of errors that you will be detecting in this step. For instance, if your QA measures are weak, you may have many systematic errors to find in the QC step. If your QC routines are not strong either, they will leave many, if not all, issues to be identified in the Data Cleaning step – including systematic issues.



Similarly to Quality Control, whenever you notice you are picking up the same errors, you should try to add a preventive routine for that to your Quality Assurance practices. If not possible, you should add an appropriate routine to your QC process since it is most likely a systematic error.

Lastly, you should make sure you document the steps and decisions taken, including any modifications to the Quality Assurance and Quality Control practices.

Equity Considerations: Data Cleaning

Because in this step you will perform the same actions as presented in the QC section, you should apply the same equity considerations raised in there. The only difference is that you will be applying them for one-off issues. See below how the same considerations for QC can be adapted to Data Cleaning.

Make sure you have a "good reason" for all of your data cleaning steps. For example, consider you found that one of the entries for a business address is wrong because the address does not exist. If location is not important for how this data will be used, there will be no need for you to make an assumption of the right address; you could correct the field to match the standards for missing values. However, if your data must capture addresses, you could make assumptions depending on your specific situation, such as searching the business online. As you can note, unfortunately, there are no standards for what defines a "good reason"– it often varies from case to case, so you will have to use your best judgment, check with others, and document your decisions.

Verify your assumptions before making any big changes. For example, consider you found the following last name in your data: "Martins." You assume this is a typo, because it looks like "Martin" - a very common last name in the U.S. You could easily act on this assumption and "correct" the entry as part of the Cleaning step. However, "Martins" is a common Portuguese last name, and changing it to "Martin" could create an error. So, before making changes, you should find ways to verify if your assumption is correct. Some ideas are to talk to people who may be able to clarify your doubt or compare the entry with other documents from the collect step. Also, make sure you document your assumptions and changes.

Talk to subject-matter experts and whoever worked in the previous data creation steps. Even if you do not spot an issue, you can reach out to these people. Because they are familiar with the data and its context, they could spot something wrong that may look right to you (according to your own assumptions).

This page left intentionally blank

A Data Journey: Data Cleaning

Before finalizing the Create phase, a team member from *NewProgram* noticed an error in the dataset below:

Call_ID	Call_Date	Call_Time	FirstName	LastName
A1234	7/24/2020	3:09 AM	Luther	Lentine
T1411	7/24/2020	4:36 AM	Ji	Vegas
A1235	7/24/2020	9:21 AM	Vince	Fairbanks
F1109	7/24/2020	11:03 AM	Maile	Mcgrath
F1110	7/24/2020	3:19 PM	Julie	Jones
A1236	7/24/2020	5:40 PM	Jo	Reep
T1412	7/24/2020	8:57 PM	Anna	Anderson
T1413	7/24/2019	10:32 PM	Matthew	Miller

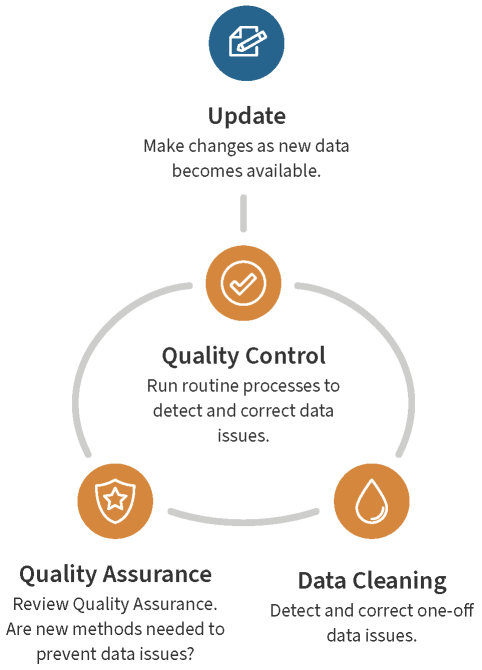
Since *NewProgram* was launched in 2020, the team member was able to quickly identify the entry as an error, but still had to investigate what caused it. They found that it was caused by a one-off system failure that temporarily modified the current year in the system used to record the entry. After ensuring that the failure did not affect any other entries, the team member corrected the call date.

Phase 2: Maintain

Most City agencies have datasets they continuously update and must maintain. **Data maintenance** is the ongoing process of detecting, correcting, verifying, and updating data entries in a database.

The purpose of this section is to give you an overview of the data maintenance process. There is one main step in the Maintain phase combined with three steps for data quality management, as presented in the image to the right. First, you will start by learning what to consider when you update your data. Next, you will revisit the concepts of Quality Assurance, Quality Control, and Data Cleaning that were discussed in the previous phase. These steps will help you ensure that your data has integrity and is useful for further analysis and reporting.

Following the Maintain phase is the Report phase.



Update

The bulk of the data maintenance phase involves updating data, which consists of adding data to an existing record, or modifying the value of data in an existing field based on new information. The new data may be collected by your agency or come from an external partner. Once you get it, you will input the new entries into the system that holds the original entries. This step will vary greatly depending on the agency and what system will be holding the data. Review the Enter step for specific considerations related to data entry.

Whoever is maintaining or updating the datasets should have the following qualifications:

- Knowledge of the system of data entry
- Understanding of the dataset
- Training in the system of entry and maintenance

Your team must evaluate resources and determine who will be doing what, including considerations for backup. You should also note that the person who is updating the dataset may not be the person who created it or owns it. Often, data gets collected by people in the field and may be handed off to other staff for maintenance.

Everyone who needs to have access to the data should be able to see it, but it is a good practice to limit who is allowed to edit and change it. Editing should be limited to those who know the system of data entry and who understand the dataset. Allowing too many people to change the data can increase errors and confusion. Conversely, too few people allowed to edit the data creates bottlenecks in the workflow process and can lead to data gaps if someone is on an extended vacation or leaves suddenly or unexpectedly.

Equity Considerations: Update

The Update step is very similar to the Enter step, with the exception that you will add to or modify an existing dataset. Therefore, you should apply the same equity considerations recommended in the Enter step: Do not let your biases enter the data and verify your assumptions. Additionally, you should follow the recommendations below:

Be aware of changes in the collection methods between updates. Sometimes, collectors will change the method they use to collect data, which could have impacts on equity. If that is the case, try to find how the new data was collected and what motivated the change.

Take into consideration that certain populations tend to be more transient. When deciding on your update frequency, consider that certain populations, such as persons experiencing homelessness, will be more transient. That is, some of their information will only be accurate and current for a short period of time, which makes keeping an updated dataset harder. In such cases, you should aim to update datasets more frequently to avoid misrepresentation.

The frequency with which data should be updated/maintained depends on the dataset. There are a few things you should think about when determining how often your data will need to be updated:

Is it a one-off (project-specific) or ongoing?
You will have to determine how often data is being entered based on need, resources, and system availability.
Will old data values be retained or overwritten?
Verify the plan for dealing with past data values when updating old datasets.
Is the dataset created by your agency or an external source?
If external, make sure you know the source's update frequency to time when you reload or update from them.
How often does new data become available?
Is it available at regular intervals, or does it have an irregular collection pattern? You will need to plan accordingly.
What other systems (internal and external) need this data?
Do you need to work with other agencies to make sure data is updated at the correct frequency for other City functions?
Is your data time-sensitive?
Do you need to immediately update your dataset for each modification in your data (e.g., a new patient address should be inputted as the new information becomes available), or can you make a group of updates at a later time (e.g., data from a quarterly survey is usually updated when the survey closes, not after each new response)?
Are reports run from this data?
Determine how up-to-date you need your data to be before you run weekly, monthly, quarterly, or annual reports, and wait until you have it.

As always, to avoid accidental loss of data, you should back up your data at regular frequencies, including when you complete your data collection activity and after you make edits to it.

A Data Journey: Update

NewProgram's datasets are regularly updated to keep patients' records current. *NewProgram*'s communications team sends out requests to patients for information updates every six months. In the current regular update cycle, two patients informed the team changes in their lives:

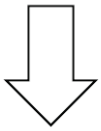
- Patient Ji Vegas got married in October, changing her last name to Davis (1).
- Patient Maile Mcgrath moved to a new address in November: 3009 Tea Berry Ln, Apt 204, Madison, WI 53151 (2).

The image below shows the update of a dataset to reflect the two changes mentioned above:

Updating Patient Information

Data on Patient Information – Original Entry in July 24th, 2020

FirstName	LastName	House#	StreetName	StreetType	Unit	City	State	ZIP
Luther	Lentine	748	Pearcy	Avenue	01	Madison	WI	53151
Ji 1	Vegas	2753	Woodlawn	Drive		Madison	WI	53085
Vince	Fairbanks	3223	Abner	Road		Madison	WI	54701
Maile	Mcgrath 2	4841	Grant View	Drive	21	New Berlin	WI	53781
Jo	Reep	4261	Reeves	Street		Milwaukee	WI	53085



Data on Patient Information – Updated in December 3rd, 2020

FirstName	LastName	House#	StreetName	StreetType	Unit	City	State	ZIP
Luther	Lentine	748	Pearcy	Avenue	01	Madison	WI	53151
Ji 1	Davis	2753	Woodlawn	Drive		Madison	WI	53085
Vince	Fairbanks	3223	Abner	Road		Madison	WI	54701
Maile	Mcgrath 2	3009	Tea Berry	Lane	204	Madison	WI	53151
Jo	Reep	4261	Reeves	Street		Milwaukee	WI	53085

Additionally, *NewProgram*'s team updates individual records every time a returning patient utilizes its services.

Quality Assurance, Quality Control, and Data Cleaning

To maintain data quality, the data maintenance process should be performed as the dataset is updated. The frequency of data quality management process depends on the update frequency. One-off projects may only need to go through the Quality Control process once since they are not frequently updated. On the other hand, continuously updated datasets may need specific quality processes that occur often, such as daily or weekly routines.

While new data should have gone through a Quality Control process, it is a good idea to perform a review to make sure everything continues to look correct. Start reviewing the methodology documentation to learn the specific quality issues you should look for in your data. Then, after the data is in your system, perform your Quality Control routines and Data Cleaning practices.

Example: Quality Management in the Maintain Phase

Once a year, the Engineering Department submits a Capacity, Management, Operation, Maintenance (CMOM) report to the Department of Natural Resources about the City's sanitary system. This report outlines the number of structures; type, length, and age of the mains; and other information.

This parcel map is regularly updated, such as when parcels are combined or split into two, or there is a larger development. When this occurs, the Engineering Department begins by defining QA measures, which include ways to ensure that all parcels are fully enclosed by parcel boundaries and that the new parcels align appropriately with adjacent parcels. It then starts to input the updated data into the CAD system.

Every other week, the Engineering Department performs a Quality Control routine before posting. The parcels are built into polygons using a script, and a change detection process shows where the changes are. Then, the data is visually reviewed and compared to the last build. If errors or inconsistencies are found, the department works on cleaning the data. After these steps are complete, the data is transformed for posting.

If you keep finding the same errors, you should review and update your Quality Assurance measures and Quality Control routines – your team will have to evaluate resources and dataset necessities to determine the best schedule for reviewing and updating QA and QC processes. Regardless of when and how you perform each step, you should always keep appropriate documentation, including all the changes made throughout the process.

Equity Considerations: Quality Assurance, Quality Control, and Data Cleaning

During the Maintain phase, you should repeat the equity considerations for Quality Control and Data Cleaning described in the Create phase. Additionally, if you need to create new QA measures, you should repeat the equity considerations for them. In summary:

Review the guidelines for equitable Quality Control and Data Cleaning. When controlling your updated dataset(s) for quality, make sure your practices align with the equity considerations discussed in the Quality Control and Data Cleaning steps in the Create phase.

Ensure that QA practices in place are not discriminatory. Now that you have real data, verify that the QA measures in place are in fact inclusive, and update them whenever necessary.

If creating or modifying QA measures, review the guidelines for equitable Quality Assurance. If you need to create or update QA measures, make sure you review the equity considerations for Quality Assurance in the Create phase.

This page left intentionally blank

A Data Journey: Quality Assurance, Quality Control, and Data Cleaning

After *NewProgram*'s team performed a scheduled regular update to maintain patients' records, it submitted the dataset to its Quality Control routine practices, which include looking for duplicated entries, missing values, and potential misspellings.

FirstName	LastName	House#	StreetName	StreetType	Unit	City	State	ZIP
Luther	Lentine	748	Pearcy	Avenue	01	Madison	WI	53151
Ji	Davis	2753	Woodlawn	Drive		Madison	WI	53085
Vince	Fairbanks	3223	Abner	Road		Madison	WI	54701
Maile	Mcgrath	3009	Tea Berry	Lane	204	Madison	WI	53151
Jo	Reep	4261	Reeves	Street		Milwaukee	WI	53085
Celestine	Zima	9627	Windsor	Drive	300	Madison	WI	
Mackenzie	Bone	521	Snake Hill	Lane		Madison	WY	53151

1

2

After applying methods to highlight duplicates in Excel, the team confirmed there were no duplicates. Then, they filtered each column to look for missing values, finding a blank zip code cell for the address of patient Celestine Zima (1). Lastly, the team also noticed a potential error in the address of patient Mackenzie Bone, as the team suspected the correct state for the address should be Wisconsin instead of Wyoming (2).

Then, the team investigated all the potential issues mentioned above to ensure they were true errors:

Issue 1 – missing zip code: After checking a web mapping platform, the team found that the zip code for 9627 Windsor Drive, Madison, Wisconsin, is 53085. The team assumed the patient forgot the zip code at the moment of response, fixed the issue, and documented the change and supporting assumptions.

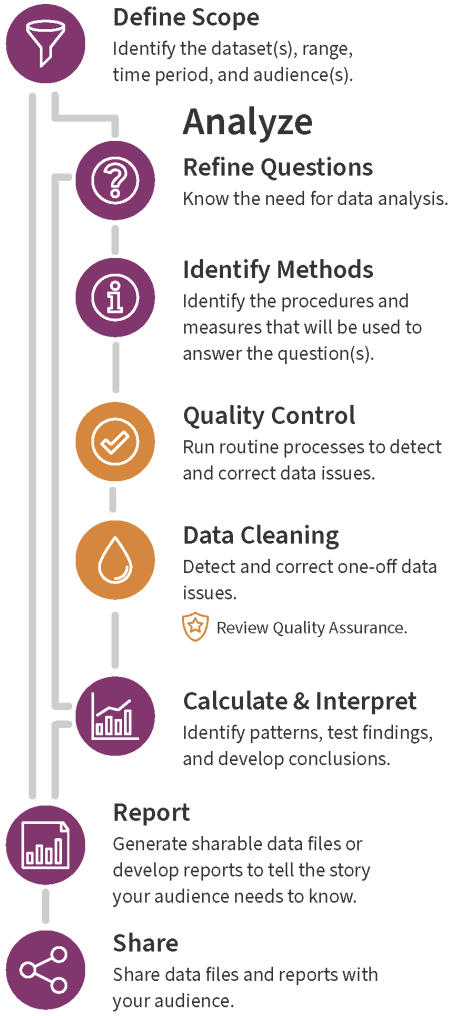
Issue 2 – potential incorrect state: First, the team verified the existence of the address 521 Snake Hill Lane, Madison, Wyoming, and found that there was no such address in Wyoming. Then, the team verified the existence of such an address in Wisconsin and found a perfect match. Because the state field uses a dropdown list, and WY is right below WI, the team assumed that the collector accidentally selected the wrong value. Next, the team corrected the state value and documented the change and supporting assumptions.

Note that at this moment, none of the issues prompted changes in the previously established Quality Assurance practices. Additionally, the team may still perform ad hoc Data Cleaning actions if other issues are flagged out of the Quality Control routine.

Phase 3: Report

Reporting and analysis take data and turn it into useful information for decision-making. In the spirit of the City's data-informed culture of inquiry, and the City's commitment to transparency, you should make ongoing efforts to report as much data as possible while protecting private and protected information.

In this final phase of the framework, you will learn the steps that lead to the reporting and sharing of data and related analyses. The Report phase has a total of eight steps, as presented in the image to the right, with two main paths. The first and shorter path encompasses datasets that can be reported directly and require a user simply publishing their data as-is. The second and longer path encompasses datasets that need to be analyzed before they can be reported. Regardless of the path you may follow, you will begin by defining the scope of your report, which is crucial for a strong groundwork.



If your scope includes data analysis, you should follow the steps within the Analyze path. In this longer path, you will start by refining questions to identify the need for data analysis. Then, you will identify the methods that will be used to answer the previously refined questions. Next, you will revisit the concepts of Quality Assurance, Quality Control, and Data Cleaning that were discussed in the Create phase. After ensuring the quality of your data, you may start your calculations and interpretations. At this point, you may find the need to circle back to previous steps in the Analyze path to ensure all questions are answered. You may also have to work concurrently in two or more steps according to your individual data analysis needs and work context. When you fully complete the Calculate & Interpret step, you may move on to the final steps of developing reports and sharing your findings.

Throughout this phase, you should collaborate and include stakeholders in the way that works best for your data project, so you can make sure you are producing reports and analyses that are comprehensive and useful.

Define Scope

Before you begin preparing data files or starting an analysis to develop a report, you need to define the scope of your project, which includes determining the project goals and making practical decisions about your reporting process. This step is crucial to ensure your data work efforts are efficiently directed to what matters and that your report will be truly useful to targeted audiences.

Equity Considerations: Define Scope

As the first step of the Report phase, Define Scope is the foundation for all the other steps in this phase. So, you must develop an equitable scope to ensure that your reports will be fair and unbiased.

Review some equity considerations in the Identify Need step in the Create phase. Both Identify Need and Define Scope serve as the foundation for the steps that follow them. So, most considerations from Identify Need can be adapted to this step:

- Be aware of your own biases.
- Make sure you understand the social factors related to the data that will be reported.
- Identify the groups of people who are unfairly affected by data practices.
- Assess how your reports will impact residents, especially those identified above.
- Identify the data that was not collected and its impact on residents, especially those identified above.
- Involve the community to help you define the scope of your report – when reasonable and possible.
- Be aware of sensitive topics and how they affect groups and individuals.

Make sure that what you want to investigate is inclusive and fair. Think about who and what you will include – or not include – in your analysis and why. Make sure you have unbiased and fair reasons for it. Also, be careful not to prioritize investigations that are easy to answer over more complex, inclusive questions without having reasonable grounds. For example, imagine an analysis project to decide where to prioritize street repairs. It is easy to go based on customer complaints, but then you will miss the needs of customers who do not feel comfortable reporting complaints.

Think broadly about the equity issues attached to your data. Start by understanding how and why your data was collected – does it have an equitable ground? Also, identify if it has gaps or limitations that can make your analysis biased, unfair, or exclusive. Note that you do not need to fix any issues in this step, but it is important to identify them from the beginning.

Take the first steps to ensure that your audience will easily access and understand your reports. An important step toward equity is to ensure that you share data in a way that your audience can use. Include in your scope broad ideas on how to ensure the accessibility and understandability of your report. For example, if one of your audiences prefers to read in Spanish, you should include “create the report in both English and Spanish” as a project need.

Most practical decisions depend on the dataset(s) that will be reported and the audience(s) that will utilize the information to be shared. Here are several key questions you should consider when defining your scope:

- Why are you reporting and sharing a specific raw file or an analysis?
- What do you want to investigate?
- What will be the purpose of your report?
- What data and contexts are you working with?
- Which dataset(s) contain the information you are seeking?
- What steps do you need to take to access the dataset(s) identified above?
- What is the range of data that is required to conduct your investigation?
- What is the time period you wish to consider in your report?
- Do you even have the data that can answer your questions?
- How will you handle the limitations in your data, such as sensitive data, missing data, or data that does not directly answer your question and must be combined with other data?
- Who are the possible audiences of your report? Who will be impacted by your report?
- Which stakeholders could, and should, be directly or indirectly involved at each step in the process to ensure their needs are met? E.g., other agencies, community partners, etc.
- How will you make sure that your scope is equitable? Read the equity considerations section for further discussion on the topic.
- Are there existing reports from previous years or comparable projects that you could use to help define your scope?

Similar to the Plan step in the Create phase, you can write a project charter or similar tool to document the scope of your project and effectively communicate all of your plans and expectations to stakeholders. Having your scope in a written document can also help you be on track with what is truly important throughout the course of the project. However, it is important to note that your scope may be subject to change in the process of the Report phase due to sudden internal or external events, such as employee turnover and policy changes.

Once the project scope is outlined, you can move on to the analysis process, or you can skip directly to the Report step since not all data needs to be analyzed before being reported (e.g., Uniform Crime Reporting data).

A Data Journey: Define Scope

The Public Health Department director requested a report about *NewProgram*'s operations for internal purposes. Then, the team responsible for analyzing *NewProgram*'s data and preparing the report met to define the scope of the project.

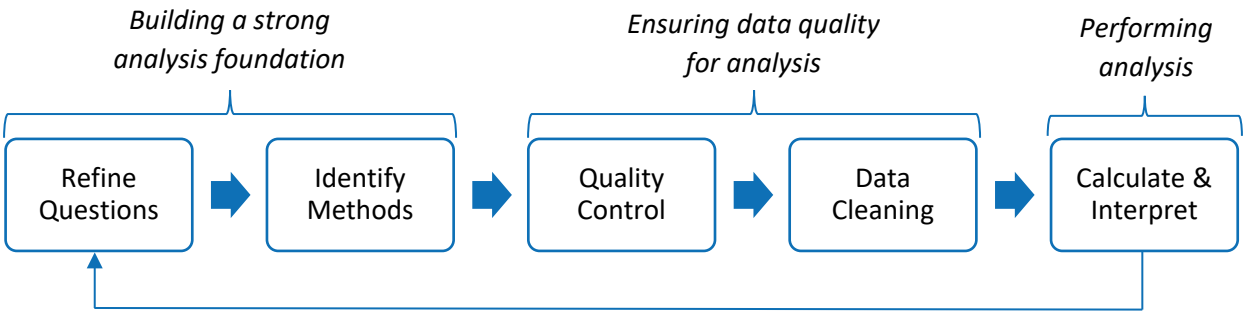
The following image shows the team's initial answers to some Define Scope guiding questions:

What is the purpose of the report?	Provide information about *NewProgram* operations
	Provide information about whether *NewProgram* is reaching its strategic goals
Which datasets contain the information needed?	Information about patients from Electronic Health Records System
	Information about the incident from Computer Aided Dispatch System
What is the range for the report?	From July 2020 to December 2020
	Only include incidents with actual response
Who are the audiences?	Public Health director
	Mayor's Office

Analyze

Data analysis is a process of systematically applying tools and techniques with the goal of discovering useful information to draw conclusions and support decision-making (Wikipedia). It consists of a detailed examination of a set of data, records, or statistics.

This section will walk you through key steps and considerations of the Analysis portion of the Report phase, from refining your questions to performing calculations and interpreting them.



Refine Questions

As previously mentioned in the Create phase, understanding the questions the data is trying to answer is a key step in defining what we will collect. In the Report phase, you may need to refine these questions to ensure proper analysis.

The following example shows how questions may evolve from the initial Create phase to the Report phase:

1. Question developed for data collection

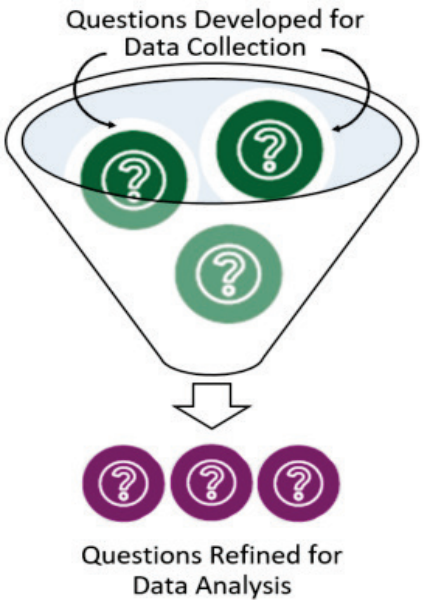
- What are the Madison Police Department calls for service?

2. Data collected to answer the questions

- Details of Public Safety Communications calls involving the Madison Police department in the last five years, including date, location, and type of call.

3. Question refined for data analysis

- How many violent crimes occurred on the north side of Madison in 2016 compared to 2017?



Note that a more general question was asked to inform data creation, and specificity was added to it in the analysis process in the Report phase.

Equity Considerations: Refine Questions

In this step, you will refine the questions developed for data collection in the Create phase into specific questions for analysis purposes. In combination with your defined scope, these questions will direct your analysis. So, it is important to make sure that they are inclusive and fair. Some points to consider are:

Find the right balance between being specific and keeping the big picture in mind. Even with an inclusive, broad scope, it can be easy to let your questions get too narrow and technical. For example, suppose your scope is to investigate pay equity. You may choose to refine your Create phase question "What is the wage of each employee at the City?" to "Is there a wage difference between genders for full-time employees?" This question is specific and follows the equitable scope you defined. However, you could be missing the bigger picture that people who suffer discrimination are more likely to have more unstable employment.

Think carefully about which groups will be included or excluded from your questions. For example, suppose that your question in the Create phase is "How many residents have diabetes?" and your scope is to understand the relationship between race/ethnicity and diabetes. Then, you refine the question as "What is the diabetes incidence rate for each racial/ethnic group?" But, what if your sample for American Indians/Alaska Natives is too small? Do you still keep them in your analysis, even though you do not have enough data to provide statistical significance? Or, do you remove them and ignore the possible health challenges this community is facing? These questions are hard to answer, and each case will have its own specificities. So, take your time to think carefully and critically before making any decisions.

Consider using qualitative data to inform your questions. When possible, seek out qualitative data from interviews, focus groups, narrative, and open-ended surveys questions. This will help you better understand the true experience of the community to develop the most relevant questions. You can also look at related existing research to see how it addressed similar issues.

Additionally, note that the specificity of your refined questions will depend on the data you have available. In the example above, you are only able to compare the years 2016 and 2017 because you have access to the dates of each call. Similarly, you will only be able to drill into a more focused geographic location, such as “the north side of Madison,” if your location data allows for this level of granularity.

Another important consideration around refining questions is that often many questions may be asked of the same dataset, including questions that were not on the radar when the dataset was first designed in the Create phase. For instance, using the example above, if time was also collected, you could refine your question to “In what time of the day does the Madison Police Department serves the most calls related to property crime?”

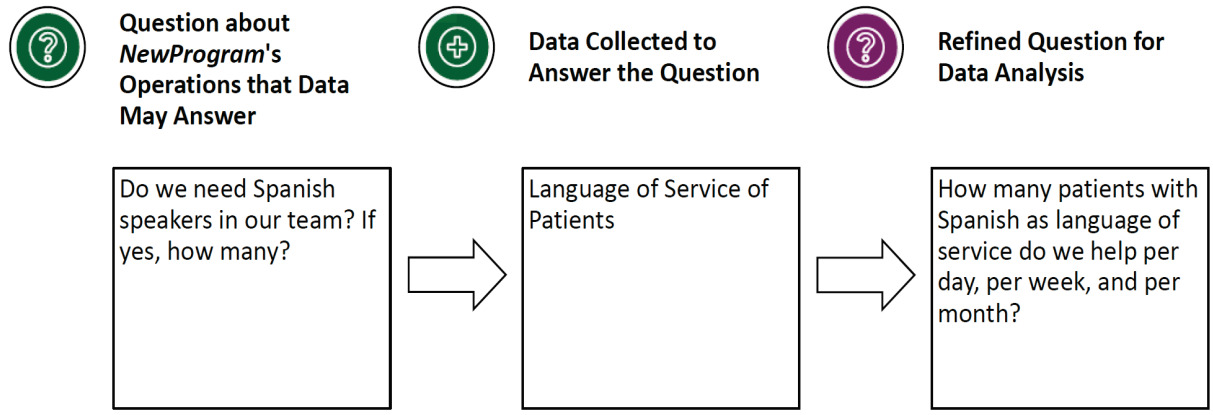
Similar to the Identify Need step in the Create phase, the involvement of stakeholders in the refinement of questions is crucial for ensuring that the right questions are being asked. So, think about which people, voices, and stakeholders should be involved and in what capacity. These could be community members, staff from other agencies, or others.

Further, as you refine your questions, consider when and where the work will be published. This and other practical decisions defined in the Define Scope step may impact and shape your questions.

Lastly, refining questions may also help further define the parameters of the project and result in updates to the project charter and/or other project documents.

A Data Journey: Refine Questions

The following image shows *NewProgram*'s process for refining a question that was developed during the Create phase to build a question more appropriate for the Report phase:



Note the difference in the relationship between the questions and the data being collected:

- The question for the Identify Need step is fundamental for identifying what data needs to be collected.
- The refined question, however, focuses on how we can analyze the data that is already collected.

Identify Methods

In this step, you will identify the procedures and methods that will be used to analyze your data and answer your question(s). This step will set the stage for a successful analysis in a similar way that adequate planning during the Create phase sets the stage for successful data gathering. Unfortunately, a "one size fits all" method does not exist. In order to choose the most appropriate method for your analysis, you will have to carefully assess the opportunities and limitations of each method you are considering and compare them to your dataset's needs and to the scope of your analysis.

Some questions that may guide you in defining the most appropriate method(s) are:

- What data will be used to answer your question(s)? Is it quantitative or qualitative in nature?
- Which category of data does your data belongs to: discrete, continuous, or nominal? See the Collect step in the Create phase for more information on categories of data.
- Does your data cover everyone or everything that you aim to study (i.e., population), or does it only cover a portion of the whole group (i.e., sample)?
- What software programs, statistical methods, or GIS techniques will be useful in helping you analyze the data? Do you have access to them?
- Are the people who will perform the analysis familiar with the method and software program of choice? Will they need training?

The table below contains some examples of data analysis methods.

<table>
<tr><th>Quantitative Methods</th><th>Qualitative Methods</th></tr>
<tr><td>Descriptive Analysis
It helps researchers find absolute numbers to summarize individual variables and find patterns. A few examples are:
• Mean: numerical average.
• Median: midpoint.
• Mode: most common value.
• Percentage: ratio as a fraction of 100.
• Frequency: number of occurrences.
• Range: highest to lowest value.</td><td>Content Analysis
One of the most common methods to analyze qualitative data.

It is used to analyze documented information in the form of texts, media, or even physical items.</td></tr>
<tr><td>Inferential Analysis
These complex analyses show the relationship between multiple variables to generalize results and make predictions. A few examples are:
• Correlation: describes the relationship between 2 variables.
• Regression: shows or predicts the relationship between 2 variables.
• Analysis of variance: tests the extent to which 2+ groups differ.</td><td>Narrative analysis
It is used to analyze content from various sources, such as interviews of respondents, observations from the field, or surveys.

It focuses on using the stories and experiences shared by people to answer the research questions.</td></tr>
<tr><td></td><td>Discourse analysis
Like narrative analysis, discourse analysis is used to analyze interactions with people.

However, it focuses on analyzing the social context in which the communication between the researcher and the respondent occurred.</td></tr>
</table>

Source: Atlan - Humans of Data.

Equity Considerations: Identify Methods

In this step, you will ensure that your methods will create fair and unbiased calculations. Here are several considerations:

Carefully choose the data that will answer your questions. Some questions are harder to answer than others. We can answer a question such as "How much fuel for buses does the city buy in a year?" by using one simple indicator: amount of fuel bought per year. But, what about a question such as "Is the new HR initiative making departments more diverse and inclusive?" To answer this and other complex questions, you will need one or more indirect sets of data. The problem is that finding this data depends on your own assumptions about the issue. For example, is the distribution of the races of employees in a department a good indicator of inclusion? It also depends on what you understand from the word "inclusion." So, it is important that you ensure your assumptions are correct and unbiased.

Weigh the benefits and risks of breaking data apart by group (especially race and gender). Breaking data apart can help you identify when a group is having worse outcomes than others. For example, median household income data broken out by race can reveal imbalances between groups. But, in some cases, it can add to the excessive monitoring of discriminated groups.

Consider when to perform intersectional analysis. Intersectionality refers to how aspects of a person's identity, such as race and gender, combine with one another, providing unique experiences. For example, the experiences of discrimination of a Black woman will be different from those of a white woman or a Black man. Intersectional analysis can help you understand these specific experiences.

Explain your methods to your audience in a way they can understand. Sophisticated methods can help you build more precise and deep analyses. But when your audience can't understand these methods, they can become a problem. Describe your methods thoroughly yet accessibly, so your audience can understand your process to easily identify inequities and make more conscious decisions.

Be careful with being too methodologically rigorous. Sometimes, when using statistics, you will not be able to show statistical significance, especially when looking at a small group. But, it does not mean the trend is not happening. In such cases, you may redefine your questions, for example, to find results that are statistically relevant. Or you can find other alternatives that highlight general trends, such as including narratives in your report.

Be aware of algorithmic bias. The use of automated decision-making algorithms is becoming increasingly common across governments. For example, some judicial systems use algorithms to decide when to grant parole to offenders. They use data about the offender (age, criminal history, etc.) to calculate the risk of re-offence. Unfortunately, there is no such thing as an unbiased algorithm, because like data, they reflect the biases of those who create them. So, for example, even if they do not use race as an explicit factor, these algorithms could still hold the pattern of prejudice against some racial groups. Location data, in particular, is often used as a proxy for race.

Other important considerations (and cautions) to have in mind throughout this step are:

- **Statistics can give you a false sense of confidence:** Sophisticated statistical methods and numbers with many decimal places can make your report seems more credible, but they do not guarantee that your numbers are a true representation of reality. Regardless of your method's complexity, you should maintain some healthy skepticism throughout the analysis process.
  For example, ask yourself: Are all the calculations correct? Are the underlying assumptions fair? Is the interpretation accurate? Is the data used for this analysis high-quality data? Does the result make sense when compared to what is understood of the context around the data? Is the overall analysis equitable?
- **Sophisticated methods are powerful, but they should not be a stumbling block to your analysis:** Sometimes, sophisticated methods can be really helpful for pointing out trends you might not otherwise see. Other times, it is better to use a simple method, either because it is more appropriate or it is what the people on your project are capable of. **Getting some understanding, as long as it is done with appropriate methods, is better than none.**

A Data Journey: Identify Methods

NewProgram's team wants to identify which areas of the city have higher demands for the program in order to allocate its resources in an equitable way. The following image shows the team's process for identifying methods and tools:

1) Questions defined in the Refine Questions step:

a) Which areas of the city have the most incidents?
b) How many incidents per service district?

2) Data used to answer the refined questions:

Call_ID	Call_Location	Call_District
A##### T##### F#####	Latitude and Longitude Coordinates	District number: 1 to 20

3) Preferred Method for Analysis should:

- Help us find call patterns for locations and districts and provide the number of occurrences per district
- Allow the creation of visuals that are simple and easy to understand (for all audiences)

4) Chosen method:

- Descriptive Analysis through visuals such as:
  - Bar graphs
  - Heat maps

5) Preferred Tool / Software should:

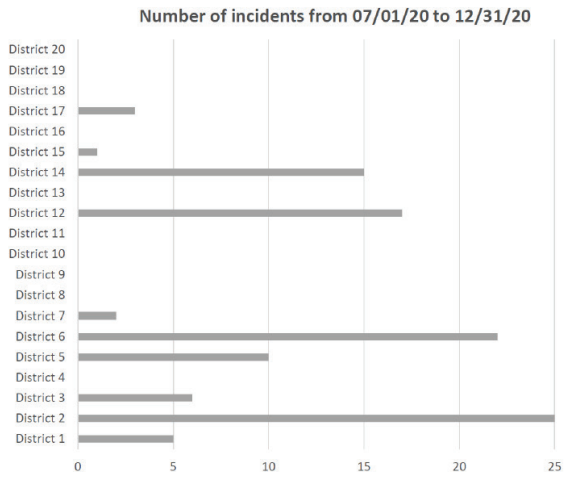
- Be able to generate visuals, such as bar graphs and heat maps
- Allow audience interaction through filters and, for heat maps, zoom in/out.

6) Chosen Tool / Software:

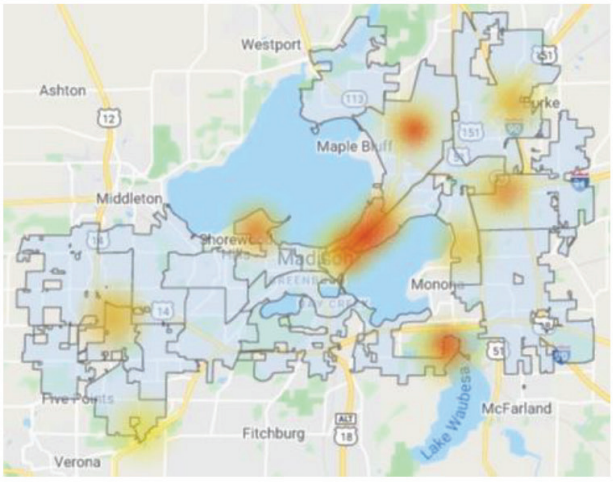
- Power BI: a tool that provides interactive visualizations and business intelligence capabilities with an interface simple enough for end users.

The images below are examples of the visuals chosen by *NewProgram*'s team:

Bar Chart for Identifying Incidents per District



Heat Map for Identifying Areas of High Incidence



Note: For simplicity, this example only uses two refined questions as the basis for the analysis. In real life, however, most analysis will have multiple questions to be answered, which can lead to multiple methods and tools.

Quality Assurance, Quality Control and Data Cleaning

While your data should have gone through a Data Quality Management process in the Create and Maintain phases, it is a good practice to review and recheck the quality of your data before starting the calculations for your analysis. Similarly to the Data Quality Management cycle in the Maintain phase, you should start by running your routine Quality Control processes and performing any necessary ad hoc Data Cleaning actions to address errors and inconsistencies.

However, on top of fixing errors and inconsistencies, in this phase, you will take additional actions in your Quality Management steps: verifying what needs to be adjusted in your dataset and modifying it for the analysis. That is, you will prepare your data not only to ensure its quality but also its usability in calculations. For example, if your analysis involves placing markers on a map based on street addresses, you may need to convert these addresses into geographic coordinates (latitude and longitude). Or, if you have daily payroll data but you are looking at pay by the week, you will need to identify groups of pay records belonging to the same person and week and sum them.

Examples of Data Quality Management Actions in the Report Phase

1. Fixing errors and inconsistencies that remained from Create or Maintain phases
   Correcting the misspelling of a street name.

Street_Name	Street_Type
Hally	Way

→

Street_Name	Street_Type
Hall**e**y	Way

   "Hally Way" corrected to "Halley Way."

2. Transforming data for analysis
   Geocoding addresses to plot a map.

Street_No	Street_Name	Street_Type
5905	Halley	Way

+

Latitude	Longitude
43.087639	-89.281189

   5905 Halley Way has a latitude of 43.087639 and a longitude of -89.281189. Two columns will be added to the dataset to include this transformation.

3. Fixing errors and inconsistencies resulting from transforming data for analysis
   Correcting an Excel formula.

Street_Name	Street_Type
Way	Way
=VLOOKUP(A2,'Addresses'!A:L,**8**,FALSE)	VLOOKUP(A2,'Addresses'!A:L,8,FALSE)

Street_Name	Street_Type
Halley	Way
=VLOOKUP(A2,'Addresses'!A:L,**7**,FALSE)	=VLOOKUP(A2,'Addresses'!A:L,8,FALSE)

   Modifying a VLOOKUP formula that is returning values from a wrong column (correcting from Street_Type, column 8, to Street_Name, column 7).

Equity Considerations: Quality Assurance, Quality Control and Data Cleaning

During the Report phase, it is common to create derived columns and rows whose values were not collected but calculated. For example, splitting a name column into first and last names, or combining daily hours worked into weekly hours worked. So, it is important to make sure your process for derivation follows the same thoughtful process as in the initial Collect step. In summary, you should:

Review the guidelines for equitable Quality Control and Data Cleaning. When controlling the quality of the data you will use in your analysis, make sure your practices align with the equity considerations discussed in the Quality Control and Data Cleaning steps in the Create phase.

Get familiar with the QA measures in place and ensure they are not discriminatory. Make sure you understand the QA measures in place and how they may have shaped the data you will use. Additionally, now that you have transformed data for analysis, verify that these measures are still inclusive. If needed, and possible, update them.

If creating or modifying QA measures, review the guidelines for equitable Quality Assurance. If you need to create or update QA measures, make sure you review the equity considerations for Quality Assurance in the Create phase.

If you identify that Quality Assurance procedures should be updated to avoid observed errors and inconsistencies, communicate that to the staff in charge of data creation and maintenance. That is particularly important for reports done regularly, so the data will not continue to present the same issues in future analyses. Still, even if you are working on a one-off project, communicating ideas for the improvement of Quality Assurance procedures is a way of sharing lessons learned with a data creation staff that may be using your project as a reference for their Quality Assurance step in another related project.

Lastly, remember that documenting changes, especially when you modify data for analysis, is critical for the creation of a replicable analysis, and it ensures that your work is transparent and easy to understand.

A Data Journey: Quality Assurance, Quality Control, and Data Cleaning

NewProgram's team started its Data Quality Management cycle by performing its Quality Control routine practices, which include looking for duplicated entries, missing values, and potential misspellings. As no issues were flagged, the team moved on to preparing the data for analysis. The team used Excel to transform the data by creating two new columns to store which day of the week (1) and which time period of the day (2) a response occurred.

Call_ID	Call_Date	Call_Time	Calculated_Day_of_Week (1)	Calculated_Time_Period (2)
A1234	7/24/2020	3:09 AM	Friday	12:00 AM - 08:00 AM
T1411	7/24/2020	4:36 AM	Friday	12:00 AM - 08:00 AM
A1235	7/24/2020	9:21 AM	Friday	08:00 AM - 04:00 PM
F1109	7/24/2020	11:03 AM	Friday	08:00 AM - 04:00 PM
F1110	7/24/2020	3:19 PM	Friday	08:00 AM - 04:00 PM
A1236	7/24/2020	5:40 PM	Friday	04:00 PM - 12:00 AM
T1412	7/24/2020	8:57 PM	Friday	04:00 PM - 12:00 AM
T1413	7/24/2020	10:32 PM	Saturday	04:00 PM - 12:00 AM
F1111	7/25/2020	12:15 AM	Saturday	12:00 AM - 08:00 AM

Note that *NewProgram*'s team used the names of the columns to indicate which columns are calculated and kept the formulas preserved to serve as documentation of how the values were calculated.

Call_ID	Call_Date	Call_Time	Calculated_Day_of_Week	Calculated_Time_Period
A1234	7/24/2020	3:09 AM	=TEXT(B3,"dddd")	=IF(C2<TIME(8,0,0),"12:00 AM - 08:00 AM",IF(C2<TIME(16,0,0),"08:00 AM - 04:00 PM","04:00 PM - 12:00 AM"))

Additionally, the team noticed an error right after creating the two new columns: the formula used for the calculation of the day of the week was referring to the wrong Call_Date cells, which resulted in a wrong day of the week for call T1413 (i.e., Saturday instead of Friday). As a quick ad hoc data cleaning procedure, the team fixed the formula to result in the correct values for the days of the week (3).

Call_ID	Call_Date	Call_Time	Calculated_Day_of_Week	Calculated_Time_Period
A1234	7/24/2020	3:09 AM	Friday	12:00 AM - 08:00 AM
T1411	7/24/2020	4:36 AM	Friday	12:00 AM - 08:00 AM
A1235	7/24/2020	9:21 AM	Friday	08:00 AM - 04:00 PM
F1109	7/24/2020	11:03 AM	Friday	08:00 AM - 04:00 PM
F1110	7/24/2020	3:19 PM	Friday	08:00 AM - 04:00 PM
A1236	7/24/2020	5:40 PM	Friday	04:00 PM - 12:00 AM
T1412	7/24/2020	8:57 PM	Friday	04:00 PM - 12:00 AM
T1413	7/24/2020	10:32 PM	Friday (3)	04:00 PM - 12:00 AM
F1111	7/25/2020	12:15 AM	Saturday	12:00 AM - 08:00 AM

Calculate & Interpret

Now it's time to determine what your data is telling you by completing your analysis with the previously defined dataset(s) and method(s) to answer the question(s) being asked. In this step, you will identify patterns and trends, test your findings, and develop conclusions.

Note that this step encompasses two actions, Calculate and Interpret, because they are inextricably linked – when you perform a calculation, you must then perform interpretation to understand what it means. And often, interpretation leads to more questions, which leads to more calculation. Sometimes your calculations and interpretations may even generate questions that can take you back to the Refine Questions step or another early reporting step.

Calculation includes:

- Manual or visual analysis – When you manually or visually count data.
  For example, counting the number of computers that have a specific application, or the number of answers to an open-ended survey question that contained comments about a specific topic.
- Computer-assisted analysis – When you use formulas and visualizations to assist your analysis.
  For example, using Excel to sum multiple expenses that happened in a year, or using PowerBI to create a visualization to show the race/ethnicity distribution of employees.
- Computational analysis – When you use sophisticated calculation methods.
  For example, running a regression in R, a programming language, to predict the relationship between two variables, or using a Natural Language Processing algorithm to group similar narratives.

Interpretation includes:

- Linking the numerical results of your calculations to real-world meaning.
- In a simple case, this may mean determining basic information about your data, such as the number of times something happened.
  For example, in the past year, we spent $100,000 on a certain activity.
- In a complex case, this may mean going beyond reporting on numbers to talking about what those numbers represent in the real world and what the implications are in a way that an audience can understand and use to take action.
  For example, if X's rate is 2 times that of Y, does that mean X is working well? Should we do more of X? Etc.

Two important points to remember here are the considerations and cautions discussed in the Identify Methods step:

- Statistics can give you a false sense of confidence, which may lead you to misinterpret your findings.
- Sophisticated methods are powerful, but they should not be a stumbling block to your analysis, and they may even lead you to more questions than answers in certain cases.

After completing these two actions, you should cross-check the results with your data to ensure they are reasonable and accurate. Additionally, you need to make sure that you have documented all your steps, including the assumptions you had to make and the limitations of your analysis.

Equity Considerations: Calculate & Interpret

This is often the most "inner-circled" part of an analysis project. That is, calculations and interpretations are often handled by a small team of data analysts – or even just one analyst. **If you are part of this inner circle, you will be making big decisions about how to understand the story behind your data, which will affect how everyone else understands and uses it**. So, you need to ensure that those decisions are transparent, intentional, and unbiased.

Recognize that interpretation is a subjective process. The numbers in your calculations are the objects of your interpretation. They carry assumptions, sometimes biased assumptions, from everyone who worked with the data up to this point. They reflect our perspectives and what we believe should be counted. Additionally, they are not the only element you will use to make sense of the story behind the data. You will also use your own experience and knowledge to draw conclusions.

Identify, check, and name your assumptions. Assumptions are a normal part of the analysis process. They help us to start making sense of the data. But some of assumptions, especially when bias is involved, can also lead to incorrect interpretations. For example, is a reduction in the number of out-of-school suspensions good or bad? At first look, you can assume that fewer out-of-school suspension is a good sign. But if this is a result of teachers feeling discouraged from reporting serious problems because it makes the school look bad, your assumption will be incorrect. So, always identify and check your assumptions. Then, make sure you document them – that way if people do not agree, they can see how that might change the conclusion.

Be aware of false positives and false negatives. These are two common risks we take when using a model to make predictions. Analyses can incorrectly identify something as true (false positive) or as false (false negative). For example, courts cannot always be 100% sure when it comes to knowing who committed a crime. If an innocent person is punished, it is a false positive case. If a guilty person is not, then we have a false negative case. In some situations, these false results will not have a significant impact, and you will be able to accept the risk. But, often these errors will create or increase social injustices.

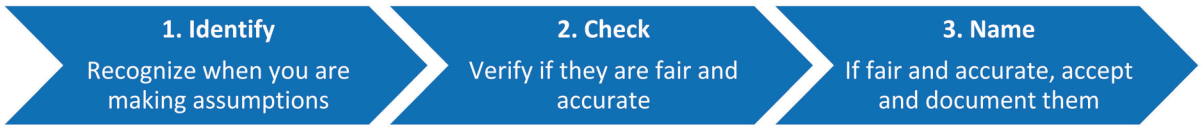
Ask a broad variety of people for their interpretation of the data. A way to reduce the chances of your biases shaping the data is to seek out the opinion of a broad variety of people. For example, you can reach out to coworkers who experience the context of the data (e.g., subject-matter experts, practitioners, data collectors). You can also involve the community in this process – if possible and appropriate.

Finally, have someone else check your results and documentation. This can range from having a teammate double-check your calculations to meetings with subject-matter experts and stakeholders to get their feedback on whether your findings seem reasonable and what their implications may be.

Sidebar 7: *Assumptions – Identify, Check, Name.*

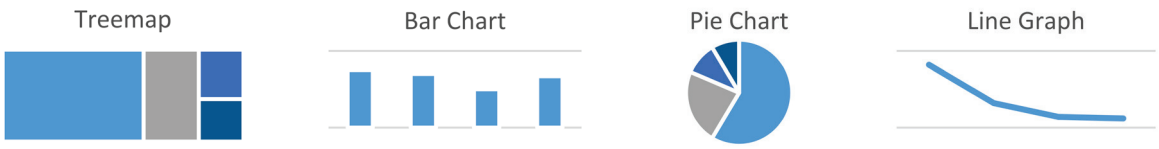
Making assumptions is a normal part of the analysis process. In fact, sometimes, you will not be able to continue your analysis without assuming certain aspects about it. Assumptions can make calculations possible and help you make sense of your results. However, you need to make sure that your assumptions are fair and add them to your documentation.

Always follow the steps below when making assumptions:



Sidebar 8: *Data Visualization.*

Data visualization refers to the graphic representation of data. Data visuals help you understand your data, and they provide an efficient way to communicate the story you want to tell.



In the *Calculate & Interpret step*, data visualization can be used to help you identify patterns and trends in your dataset. The visuals developed in this step can serve as the first analysis attempts and lead to more elaborated questions and calculations, as well as being used to check and interpret calculation results. In the *Report step*, these visuals will be polished to be shared with your audience.

Either way, it is important that you:

- **Find the best visual to use**. Data Viz Project is a great tool that provides a dictionary of different types of visuals that could be used to communicate data comparisons, correlations, distributions, geospatial outcomes, and trends over time.
- **Use your design elements wisely**. Consider the many different design elements, such as color, size, shape, opacity, texture, position, and orientation. For instance, the colors used in a visual may have an important impact on how you and your audience interpret your data. The use of warm colors like red or orange often indicates loss or danger. Other colors, like blue or green, can indicate growth or positive outcomes.
- **Get familiar with tools that allow for data visualization**. There are many tools to choose from. Some common options at the City are Excel, PowerBI, and ArcGIS. The Data & Innovation Team has a PowerBI Learning Resources page for employees interested in learning the tool.

Source: Apolitical

A Data Journey: Calculate & Interpret

After identifying the appropriate methods to answer the refine questions, *NewProgram*'s team started the Calculate & Interpret step. The following examples show two calculation methods used by the team and their interpretations:

Example 1: Count of *NewProgram*'s Incidents

1. **Question:** How many *NewProgram's* calls were there from 07/01/2020 and 12/31/2020?
2. **Calculation:** Excel Formulas (computer-assisted analysis). In this example, we could use the formula COUNTIFS on excel to count the number of incidents that happened between 07/01/2020 and 12/31/2020.

Call_ID	Call_Date
A1280	07/01/2020
T1479	07/01/2020
F1070	07/01/2020
⬇	
F1095	12/31/2020
T1504	12/31/2020
A1334	12/31/2020

3. **Interpretation:** *NewProgram* responded to X calls between 07/01/2020 and 12/31/2020.
   This is a simple case of interpretation, where the number of times something happened is really what we care about.

Example 2: Comparison of Patient Outcomes: NewProgram vs. OldProgram

1. **Question:** Do patients in *NewProgram* have better outcomes than patients in *OldProgram*?
2. **Calculation:** Linear regression analysis (a powerful statistical method that allows you to examine the relationship between two or more variables of interest). The table below shows the results of the regression analysis for *OldProgram*'s and *NewProgram*'s rates of appropriate care provided during and after a call response.

	Care Provided			Follow Up Provided		
	Rate	95% CI	*p*	Rate	95% CI	*p*
OldProgram	reference			reference		
NewProgram	2.1	1.5-3.0	<0.01	1.7	1.2-2.4	<0.01

3. **Interpretation:** Based on the table above, NewProgram's team can conclude that (1) The rate of *NewProgram*'s patients who receive appropriate care is 2.1 points higher than the rate of *OldProgram*'s patients, and (2) the rate of *NewProgram*'s patients receiving follow-up care is 1.7 points higher than the rate of *OldProgram*'s patients. Thus, their interpretation might be stated as "*NewProgram* provides better outcomes to patients than *OldProgram*."

Report

Now that you are ready to report, it is time to determine the best way to present the information to your identified audience. Is this report produced internally for the agency? For other City employees? Is it a special request from a researcher or resident? No matter the audience, you should consider the following points before creating your reports:

Define the story you are telling your audience. What are the most important points you want your audience to walk away with? How can you share these points accurately, with the necessary nuance, yet in a way your audience can understand and use to take action? How can you ensure equity when presenting the story? And finally, how adept is your audience in understanding your data and analysis?

Consider how the information is presented. Sometimes, this is simple – you might simply pull together a shareable data file and pass it on. Sometimes, this is more complex – you might have conducted an analysis, and now you must tell the story and interpretation your audience needs to know, based on the results of your analysis. This more complex presentation could be a full report with visuals and extensive written components, like the Alcohol Study (2019), a simple one-pager that provides a quick answer to a question, a dashboard to allow your audience to interactively explore data, or slides for an oral presentation.

Additionally, consider what design elements and visuals will be used and how accessible your report is to the reader. Think of practical questions, like how a person with a visual impairment would consume the report, or if the report could be made available in other languages.

Lastly, keep in mind that data visualization can be used as a tool to provide clear and accurate information, but it can also cause misunderstanding if used improperly. Thus, always be careful when choosing your visuals and design elements to avoid providing misleading information.

Include selected parts of your documentation and contextual notes. Not all of your documentation needs to, or even should, be included in your report. However, certain limitations, assumptions, and particularities of your calculations should be made available to avoid misinterpretation. Additionally, you may need to include contextual notes to either provide additional qualitative information or move the audience away from common misinterpretations (e.g., confusion between correlation and causation).

Be aware of the local, state, and federal laws that protect privacy. Regardless of whom the report is formally addressed, you will need to determine if there are privacy concerns before developing it. As previously mentioned in the Data Privacy section, there are three broad privacy classifications:

> **Public** - This data can be publicly disseminated without any concerns.
>
> **Protected** - This data is protected by law or regulation and can only be shared or accessed by a limited group or through a limiting procedure; if cleaned to remove certain information, or aggregated, this data could potentially be shared.
>
> **Sensitive** - This data is not regulated like protected data, but in its raw form, this data poses security concerns and could potentially target individuals or pose other concerns; if cleaned to remove certain information, or aggregated, this data could possibly be shared.

Equity Considerations: Report

Even if you have carefully considered equity throughout your analysis, your report can still reflect your own biases. On top of that, people can misuse your findings to maintain or increase unfairness. So, in this step, you should consider how to prevent bias and misinterpretation. Additionally, you need to make your report accessible while respecting privacy limitations.

Ensure your audience can access and understand the content in your report. People can make informed decisions when they have information and can understand it. Think about how you can make sure your reports are easy to access and understand. Some examples: use plain language, translate materials, and choose color palettes suitable for color blindness.
Check the City's Content, Accessibility & Plain Language Tip Sheet.

Use inclusive language in your report. Careless word choices can exclude groups and individuals who already face discrimination. For example, if you label "transgender woman/man" separately from "woman/man" in a graph, you will be reinforcing the harmful assumption that cisgender is the norm.
Check the City's Gender-Inclusive Language Style Guide

Be aware of how your visualizations and words can influence people's interpretations. For example, warm colors like red are often used to indicate loss or danger. Similarly, words such as "alarming" and "rising" send the message of urgency. But there are other choices that are more subtle yet harmful. For instance, the order of race/ethnicity labels in a graph can reinforce harmful world views. That is, if you put white before all other races without any clear reason, you could be sending the message that white is the norm or the priority.

Include qualitative stories to contextualize quantitative data – when possible and appropriate. Numbers alone can be striking but they don't necessarily tell a story. For example, consider that you reported a higher incarceration rate for racial minorities. A person could use this rate to support unfair arguments, such as "racial minorities commit more crimes." In this case, you can prevent the misuse of data by including a narrative that explains the impact of racism on the rate.

Address data privacy and re-identification risks. When creating your report, you should apply measures to control the risk for data privacy issues. For example, some reports that include data protected under HIPAA, such as DHS sexually transmitted disease surveillance report (2018) – Dane County, will not show data for diseases with less than five cases reported. But remember that your audience can still combine your report with other public documents to identify someone. So, make sure you add appropriate measures to control for re-identification risks.

Be transparent about your analysis and reporting choices. Make sure you document the reasons behind your choices, especially if you made a choice that could create equity issues, even if it follows external requirements. For example, consider that you analyzed the wage gap between races. Because of the size of your samples, you had to combine multiple races together to make statistical comparisons (e.g., white compared to all other races grouped together). Although your choice improved the certainty of your analysis, it also increased the danger for equity issues, such as reinforcing the idea of white as the normal and centric racial identity. You could balance this issue by explaining that this is not how the City of Madison perceives the world and providing your reasons for the choice.

When possible and appropriate, seek feedback from a select group of stakeholders before sharing your final version. Obtaining feedback from a select group of stakeholders, such as a steering group or project sponsors, can help you identify if your draft is understandable and aligned with the objectives defined in the project scope. This is especially important when reporting analyses that involve sophisticated calculations, complex interpretations, and/or difficult topics.

Sidebar 9: *Storytelling with Data Visuals.*

Data visualization is a form of storytelling. Data visuals can communicate simple, clear, and visually engaging stories about your data. In order to create stories through data visualization, you need to develop your visuals using the right elements (read *Sidebar 8: Data Visualization*) and a compelling narrative.

See below for some best practices for data storytelling:

- **Remember your audience**. Target a specific audience; ask yourself what is important to your audience; and decide on a couple of key questions to answer.
- **Work with clean data**. Do not start by doing anything fancy – get your data ready first; make sure your data is not confusing or full of errors; and free your data from duplications.
- **Stick to the story**. Focus on developing a main trend and not getting distracted by other findings; ensure your story is clear – clarity is often what makes stories really good; and stick to a few clear takeaways and explain them clearly for the reader.
- **Humanize the data**. Try to include real human stories that showed the lives affected behind the data – numbers alone can be striking, but they do not necessarily tell a story.

Additionally, you should submit your visuals and story to an equity analysis process, such as RESJI's Analysis Tools, to ensure equitable outcomes and reduce unintended consequences.

Source: How to think like a data journalist.

A Data Journey: Report

After performing all calculations and interpretations, *NewProgram*'s team finalized the detailed report requested by the Public Health Director. The image below is a screenshot of page 17 of the written report:

Previous research identified that patients who receive appropriate care as part of the emergency response and return within a week for a follow-up appointment present better outcomes. Thus, the number of in-call cares and follow-ups provided in each program were used as indicators to respond to the research question "*do patients in NewProgram have better outcomes than patients in OldProgram?*"

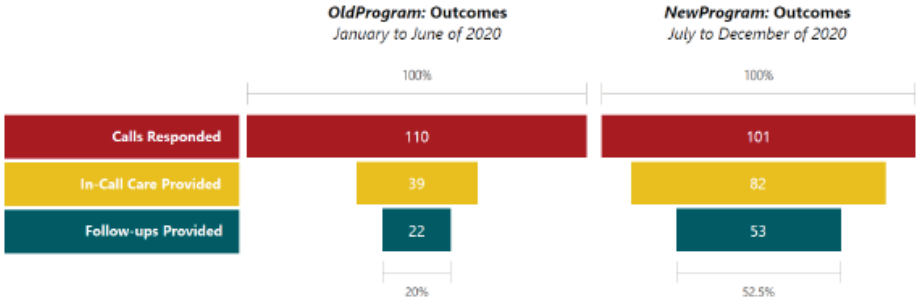


Figure 7: Comparison of Patient Outcomes between OldProgram and NewProgram.

The figure above shows that only 1 in 5 patients who were served by *OldProgram* returned for their follow-up appointment. Under *NewProgram*'s operations, however, 1 in 2 patients completed their program of care.

Regression Analysis

To further confirm that *NewProgram*'s patients have better outcomes than *OldProgram*'s patients, the project team examined the statistical relationship between in-call cares and follow-ups provided in *OldProgram* and *NewProgram*. The team identified linear regression analysis as the appropriate statistical method for this study.

Table 11, below, shows the results of the regression analysis for *OldProgram*'s and *NewProgram*'s rates of appropriate care provided during and after a call response.

	Care Provided			Follow-up Provided		
	Rate	95% CI	*p*	Rate	95% CI	*p*
OldProgram	reference			reference		
NewProgram	2.1	1.5-3.0	<0.01	1.7	1.2-2.4	<0.01

These findings can be interpreted to mean that:

- The rate of *NewProgram* patients who receive appropriate care is 2.1 points higher than the rate of *OldProgram* patients who receive appropriate care.
- The rate of *NewProgram* patients receiving follow-up care is 1.7 points higher than the rate of *OldProgram* patients receiving follow-up care.

Note that multiple elements (graphs, tables, text) were used to present one of the elements of the story: *NewProgram* provides better outcomes to patients than *OldProgram*.

Share

This final step encompasses the process of sharing and publishing data files and reports to previously identified audiences. It can be accomplished through a wide range of methods, including but not limited to:

Publishing a dataset to the City's Open Data Portal

Publishing a dashboard to the web

Sharing a written report via email

Delivering an oral presentation

However you do it, this step is where your previously identified audience learns about your work, your data, and, where appropriate, your interpretation of it.

The City of Madison aspires to openly share the data it holds in public trust where appropriate (Madison General Ordinances 3.72 – Public Accessibility to Municipal Datasets). However, various local, state and federal laws protect privacy. Thus, not everything can or should be shared in such a way.

DataSF, the data team at the City of San Francisco, notes the publication of data requires balancing several factors, including the value of having that data available publicly versus the likelihood and associated risks of someone or something being identified through the data (DataSF). You should consider the following questions in determining whether the data and report can be shared on the City's Open Data Portal:

Is there information protected by law or regulation (such as health information protected by HIPAA) in the data?

Is there data that could identify a specific person in the data (such as birthday, address, and agency of employment)?

Is there data in that dataset that could, in combination with other publicly-available data, identify an individual?

Equity Considerations: Share

Sharing data is an important action for increasing equity. When we share reports built with equitable roots, we increase fairness in decision-making. When we publicly publish files, we make our decisions more transparent. But, it is important to take the considerations below into account:

Address data privacy and re-identification risks. Like in the Report step, you should apply measures to protect the privacy of people included in your data files and reports. Start by identifying which privacy category (public, protected, or sensitive) they belong to. But remember that your audience can still combine them with other public documents to identify someone. This risk increases for individuals who belong to small minority groups. For example, if you release salary data without names, but with demographic data, there may be many white men, but only one female Native American, so her salary can be easily found while her colleagues' cannot.

Weigh the benefits and risks of publicly sharing data files and reports. In some cases, publicly sharing data can add to the excessive monitoring of discriminated groups. But not releasing data that is needed to understand and address community issues may also lead to unfairness. Also, some groups are more targeted for data collection and studies. So, when you share data, you also reduce their burden of having multiple organizations seeking the same information from them.

Ensure community members can easily access and use shared data. We have the power to make data truly useful by ensuring that our audience knows how to access and use our reports. Some actions you can take are: publishing your reports to the City's open data portal, being mindful of how hard it is for your audience to find your report, partnering with organizations that work with diverse groups, and providing action-oriented recommendations for next steps.

Note from these questions that it is important to consider both whether the data itself presents a risk and if the data could be used in combination with other publicly-available data to create a risk.

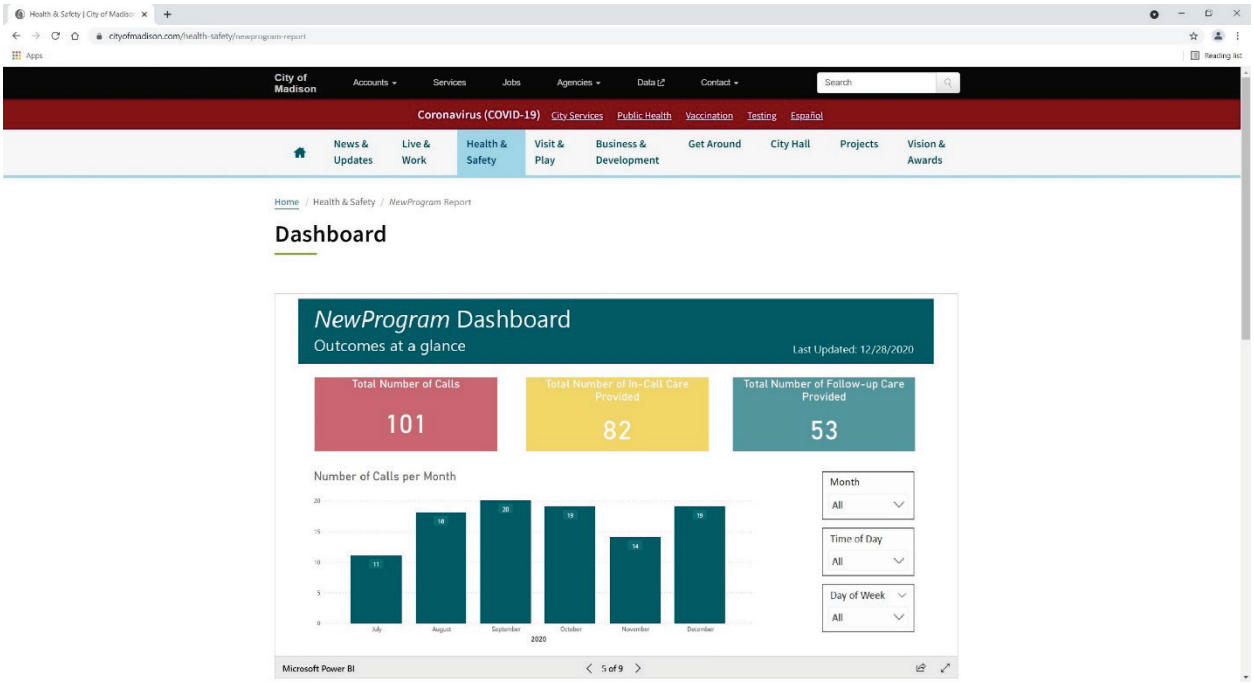
Additionally, whenever publishing data, it is important to include the appropriate metadata (see Sidebar 5: Metadata) so that others understand the source, aspects, and limitations of your data.

Although this is the last step in the Data Management Framework, you may have to adjust your report or even start the cycle over (e.g., start a completely new data project) based on new stakeholder questions motivated by the shared report.

Thank you for your commitment to following data management best practices.
This guide will be updated and expanded as needed.

A Data Journey: Share

In addition to the written report requested for internal purposes, *NewProgram*'s team simultaneously worked on building a dashboard to be shared with the general public. The image below is a screenshot of the dashboard published on the City's website:



Appendix A: Glossary of Key Terms

Term	Definition	Example
Data	Information that can be examined, considered, and used to inform decisions.	Location of bus stops; voter turnout rates; hourly water usage.
Data Analysis	A method of systematically applying tools and techniques with a goal of discovering information to support decision making.	The Alcohol Study (2019) utilized city data to analyze the density of alcohol licenses throughout Madison, and to determine if there is any relationship between alcohol license density and public service utilization.
Data Analytics	Refers to the use of data analysis methods to describe, predict, and improve organization performance or solve problems.	Creating a PivotTable in Excel to see if there is a relationship between voter turnout and proximity to a polling place.
Data Cleaning	The steps of detecting, correcting or removing inaccurate data entries in the dataset.	Correcting any misspelled street names in the labeling of Metro bus stops after a new route is added.
Data Collection	The process of collecting information from a variety of sources.	Meters record hourly water usage.
Data Documentation	Data documentation outlines what is being/has been done, how, and by whom. See also: Sidebar 4 in the Quality Assurance section.	A Word document that explains how two datasets were merged together.
Data Entry	The transcription of information into an electronic file or database.	Clerk staff enter new registered voters into WisVote.
Data Ethics	Refers to the code of behaviors and practices that helps ensure that everyone handles and uses data ethically.	Being transparent with residents about the City's data practices and use.
Data Equity	Refers to the consideration, through an equity lens, of the ways in which data is collected, analyzed, interpreted, and distributed.	Identifying how your data analysis will impact underserved communities.
Data Governance	Refers to the means by which an organization makes decisions about its information assets.	This Data Guide and the APM are tools the City of Madison uses to govern its data management.
Data Maintenance	The ongoing process of detecting, correcting, verifying, and/or updating the data entries in the database.	Running a weekly script to automatically correct misspelled street names.
Data Management	Refers to the implementation of practices to ensure the overall management of the availability, usability, integrity, and transparency of data across an organization.	City staff taking care to follow standards and procedures with their data.

Term	Definition	Example
Data Ownership	This refers to both the possession of and responsibility for information. This includes the ability to access, create, modify, package, or remove data, and also the right to assign these access privileges to others.	Human Resources owns the Neogov system and its data.
Data Point	A single piece of information.	The location of a specific bus stop.
Data Privacy	The considerations around utilizing and sharing data while protecting personal information. Data privacy has three broad categories: public data, protected data, and sensitive data.	Public data: City of Madison Neighborhood Association map. Protected data: Employee data related to the Family and Medical Leave Act (FMLA). Sensitive data: Aggregated race and ethnicity data of COVID-19 cases.
Data Quality	A perception or an assessment of data's ability to serve its purpose in a given context as reflected by factors such like accuracy, completeness, consistency, validity, and verifiability. A dataset that truly reflects reality, is appropriately filled out, does not conflict with other datasets, measures what was intended to be measure, and aligns with an existing and verifiable source.	A dataset that truly reflects reality, is appropriately filled out, does not conflict with other datasets, measures what was intended to be measure, and aligns with an existing and verifiable source.
Data Silo	A repository of fixed data that remains under the control of one agency and is isolated from the rest of the organization.	Two departments in a division keeping their own counts of the same statistic using different methods, rather than sharing information.
Data Source	Refers to where data comes from and who collects it. A data source drawn from within an organization is called an internal source, sometimes called a primary source. A data source from outside the organization is called an external source, sometimes called a secondary source.	An internal data source would be the Assessor's parcel data. An external data source used across the city is the U.S. Census Bureau.
Data Standards	The rules by which the data is described or recorded to ensure consistency and comparability across datasets.	Where all agencies follow a set date format such as "10-31-2019." See also Appendix B: City of Madison Data Standards.
Data Stewards	Employees who work to improve the handling of aggregate data and systems thorough stewardship (see Data Stewardship definition below).	Changing how the agency record information about gender of all clients.

Term	Definition	Example
Data Stewardship	The implementation of data governance policies and the oversight of data management practices within departments.	Ensuring that everyone in an agency is following the data standards presented in this guide.
Data Storage	A general term for where and how data is saved in electronic or other forms.	Information about Alcohol Licenses is stored in the Accela Permitting system.
Data Type	A classification of data which tells a computer system how to interpret its value.	In Excel, you can store a number as a number type or as a text type. Only the number data type will allow you to perform mathematical operations.
Data Visualization	Refers to the graphic representation of data. It helps you understand your data, and it provides an efficient way to communicate the story you want to tell.	A pie chart illustrating the numerical proportion of employees by gender.
Database	Any collection of data or information, which is specially organized to facilitate the storage, retrieval, modification, and deletion of data in conjunction with various data-processing operations.	The Parking Utility tracks its assets in a database.
Dataset	A collection of data that is related by content and structure.	The “voter turnout by ward” dataset is a collection of voter turnout rates.
Dataset Inventory	The compilation of information about all datasets and data sources available, such as what format that data exists in and who is the owner.	A spreadsheet listing all datasets available to an agency with details about their content, location, owner, and update frequency.
Enterprise Database	A database supporting an enterprise system.	The database supporting the MUNIS system.
Enterprise System	A large-scale application software package that supports business processes, information flows, and reporting.	The MUNIS system by the City to track spending and other transactions.
Equity	Refers to the allocation of resources and opportunities to provide equal outcomes to all residents.	DCR’s Affirmative Action Student Professionals In Residence (AASPIRE) internship, which offers underrepresented groups on-the-job experience with the City of Madison.
File Format	The layout of a file in a format recognizable to a specific program.	The file ending ".xslm" indicates this particular file is a macro-enabled Excel file. It means this data and all its features can be viewed and manipulated in Excel.

Term	Definition	Example
Machine Readable	Information that is directly usable by a computer or in a format that can be easily processed by a computer.	GIS files are machine readable. PDF files are not machine readable.
Metadata	Data that provides information about other data. See also: Sidebar 5 in the Quality Assurance section.	Your agency's dataset inventory is metadata!
Open Data	Data that can be freely used, reused and shared by anyone.	The City of Madison openly shares the data it holds in public trust on the Open Data Portal.
Service Indicator	A measure tied to an activity or service. The measure is then used to explore a line of inquiry to determine success or areas in need of improvement.	The number of registered voters divided by eligible voters by ward.
Personally Identifiable Information	Information about or pertaining to an individual in a record which could be associated with or traced to the individual.	Employee names and home addresses.
Quality Assurance (QA)	A proactive process to ensure data quality by preventing errors from entering or staying in a dataset. See also: Sidebar 6 in the Quality Control section.	Payroll clerks set standards for the entry of employee hours to ensure correct pay.
Quality Control (QC)	A system of routine, planned procedures established to measure the effectiveness of the Quality Assurance process. See also: Sidebar 6 in the Quality Control section.	Utilizing Excel tools to find duplicates in the data.
Records	Materials created or kept by an organization, containing information relating to the function of that organization.	A Word document with results of an employee investigation.
Records Custodians	Employees designated to take care of records, often working with the content of individual records.	Finding records for an open record request.
Results Madison	A component of the City's strategic framework that uses data-based indicators to help us better understand our service delivery and where to target improvements.	A group of data-based indicators per Agency's service.
Software	A collection of computer instructions that tell the computer how to work.	Adobe is a software package.

Appendix B: City of Madison Data Standards

Introduction

The goal of the City of Madison Data Standards is to make data creation, maintenance, and reporting easier for all users. This goal requires high-quality, interoperable data, in which standards play an important role. For example, standards can enable the sorting, filtering, and joining of data without requiring programming skills.

Here, you will find standards, based on industry best practices and data equity considerations, for recording and formatting data to reduce the chances of inaccuracies and inconsistencies across datasets, helping the City advance its data usage.

As actors of change, **City staff who work with data at any level are expected to engage with the City of Madison Data Standards**. These standards and practices are intended to be broad and general enough to cover many data contexts observed in the City, and data users are expected to refer to them and conduct data-related work in accordance with them.

However, it is acknowledged that these standards and practices will not fit the needs of every agency and every situation. This may be because they do not cover a specific subject area, because of external factors like mandatory reporting requirements or vendor capabilities, or other situations.

In these cases, when data users face difficulty implementing these standards, data users are expected to attempt to find solutions that are in alignment with both the APM and Guide, and the external requirements; and to use their best judgment to tailor the APM and Guide's standards to their situations.

Moreover, in these cases, data users with needs not met by the existing APM and Guide should not simply ignore these standards and practices. This leads to the very problems with data quality and siloing the APM and Guide are intended to address, and ultimately hamper the City's ability to make data-informed decisions.

Instead, data users are expected to engage with the Data Stewardship Program and the Data and Innovation Team to update the City of Madison Data Standards to reflect their needs. The APM and Guide are living documents, and in this way, we will continue to build together comprehensive data standards and practices to meet the needs of our City.

Standards

NAME OF PERSONS

Multiple fields to accommodate sorting and filtering:

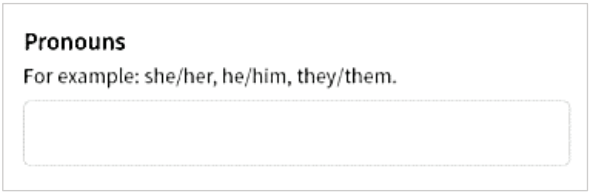
- FirstName
- MiddleName (or initial) – *may be left blank if unknown*
- LastName
- Suffix (e.g., III, IV, Jr., Sr.) – *may be left blank if unknown*

DEMOGRAPHICS

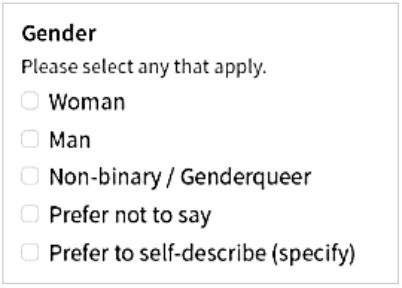
!! *If you must record gender or sexual orientation demographics, BEFORE collecting and entering data, make sure to read the Attachment 2 – Language Style Guide in the APM 2-52 Inclusive Workplace - Transgender, Gender Non-Conforming, and Non-Binary Employees, especially the section about best practices for forms & demographics* **!!**

1. **Gender & Pronouns**
   Overall, avoid collecting gender, if possible. Think critically if you need to know a person's gender or just the pronouns they use.

   Standards for Pronouns:
   Allow users to choose their own language, as seen below:



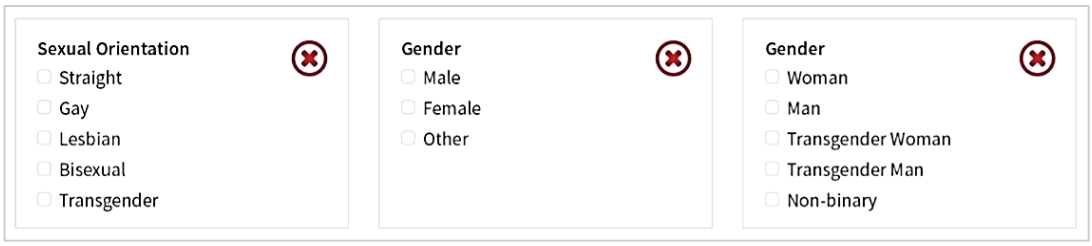
   Standards for Gender:



   ***Note**: Forms must allow users to choose multiple options, and fields must be optional (in addition to including "prefer not to say").*

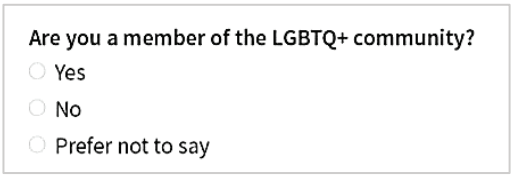
   If you must know if a person is transgender, follow the general standards for gender, and add the field below:

   Incorrect Fields:
   Do not use any of the following:
   - List Transgender under Sexual Orientation.
   - Place non-binary genders into an "Other" category.
   - Use the labels "Male" and "Female."
   - List Transgender Woman/Man separately from Women/Man.



2. **Sexual Orientation**
   Avoid collecting sexual orientation, if possible. This topic is very personal for many people, and some users will be very uncomfortable with this question.

   If you must know if a person's sexual orientation, use the question below instead:



Note: This question may be used in addition to question(s) about gender.

LOCATIONS

1. **Street Addresses**
   Multiple fields to accommodate sorting and filtering:
   - House Number
   - Street Direction – *may be left blank when not applied*
   - Street Name
   - Street Type
   - Unit Number – *may be left blank when not applied*

   Note: The above recommendations intentionally do not specify data entry conventions for each field. For example, you may use N. or North in a street direction field. However, we recommend internal consistency within agency whenever possible.

2. **City**
   One field – Always separate from other location fields, such as state.
   E.g., Correct: Madison / Incorrect: Madison, WI

3. **State**
   One field – Always separate from other location fields, such as city.
   E.g., Correct: WI / Incorrect: Madison, WI

4. **ZIP**
   One field – Always separate from other location fields.
   Can include ZIP+4 if the data is available

5. **Map Location**
   Latitude and longitude coordinates, in decimal degrees.

E.g., the latitude and longitude coordinates, in decimal degrees, for the City-County building are: 43.07244745307223, -89.38210434570038

6. **Parcel Numbers**
   Always store as text, with leading zero, and no dashes.

DATE & TIME

1. **Date**
   Use mm/dd/yyyy

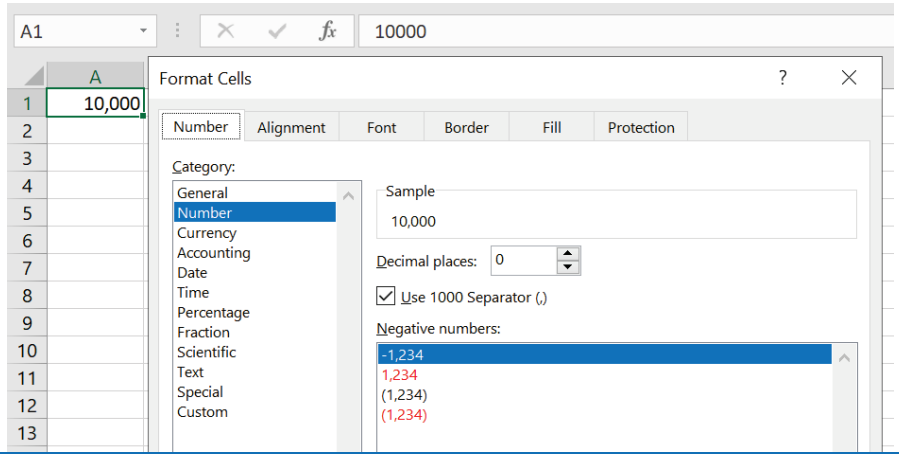
2. **Time**
   Use hh:mm:ss AM/PM

   Daylight Saving Time: If you are using a software that does not account for time changes due to daylight savings AND this is an important matter for the quality of your dataset, we recommend adding another column to store the UTC offset as "UTC±00:00".
   E.g., During daylight saving time, Madison is five hours behind Greenwich Time, so use "UTC-05:00". Otherwise, Madison is six hours behind, so use ""UTC-06:00".

NUMERIC VALUES

1. **Numbers**
   When using software that DOES NOT support fields that differentiate data types (e.g., Microsoft Word treats numbers, dates, currencies, and percentages all as characters):
   - Always store without commas.
     E.g., Correct: 10000 / Incorrect: 10,000

   When using software that does support data types (e.g., Excel):
   - Make sure you have the right data type (e.g., a number) – *the software will store the number without commas, even if it displays it in a different way.*
     E.g., if you enter "10,000" in an Excel cell, it will automatically recognize it as a number data type and store it as "10000" but keep displaying it as "10,000." However, if you enter "10000," you will have to manually change the data type.



Note: Use decimals where appropriate so the consumer does not have to know about certain specific business rules

E.g., take number and divided by 100 before using in a calculation.

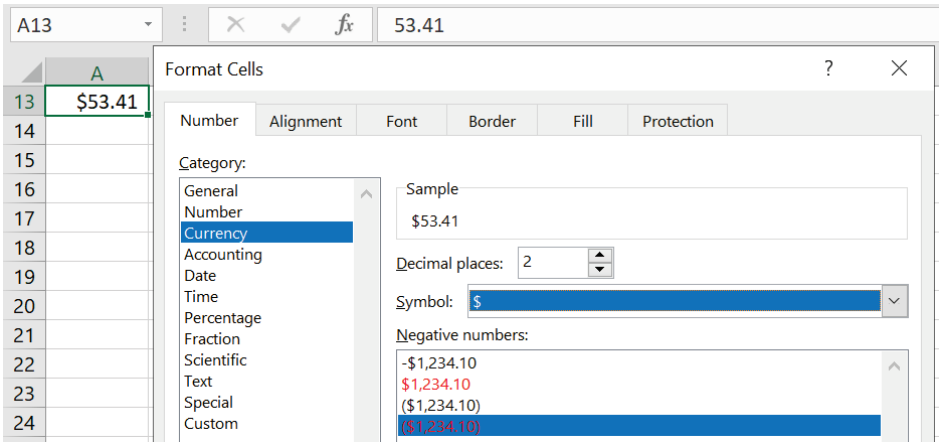
2. Dollar Amounts

When using software that DOES NOT support data types:

- Always store with at least two places to the right of the decimal.
  E.g., Correct: 53.00 / Incorrect: "53"

When using software that does support data types:

- Make sure you have the right data type and keep two places to the right of the decimal.
  E.g., if you enter "$53.41" in an Excel cell, it will automatically recognize as it as a currency data type and store it as "53.41" but keep displaying it as "$53.41." However, if you enter "53.41," you will have to manually change the data type.



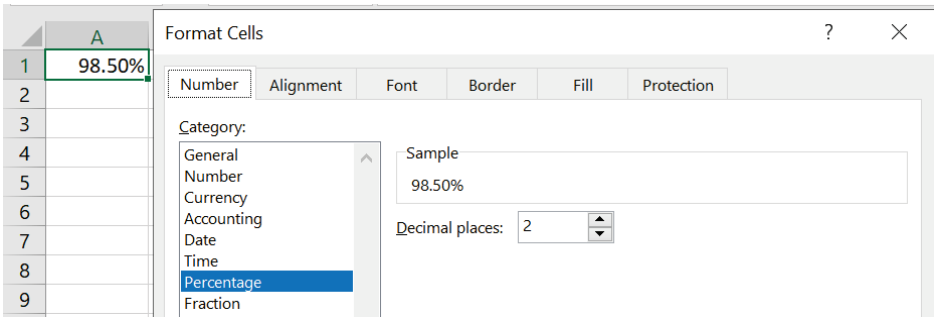
3. Percentages

When using software that DOES NOT support data types:

- Always store as decimal - *to accommodate using that value in formulas*.
  E.g., 98.5% should be stored as 0.985.

When using software that does support data types:

- Make sure you have the right data type.
  E.g., *if you enter "98.5%" in an Excel cell, it will automatically recognize as it as a percentage data type and store it as "0.985" but keep displaying it as "98.5%" However, if you enter "0.985," you will have to manually change the data type.*



4. Measurements

Make it clear what the units of measurement are by either:

- Including it in the column name – *if all measurements are in the same unit.*
  E.g.,

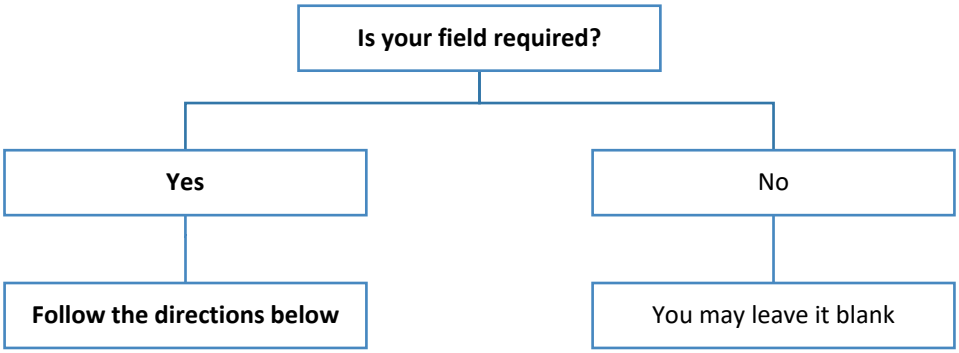
Weight_lbs
140.2
121.5

- Or, making a separate column with each measurement's unit of measure – *If the units of measure may vary*
  E.g.,

Weight	Unit
140.2	lbs
55.1	kg

Note: *converting everything in the columns to the same unit of measure is strongly recommended.*

MISSING DATA FOR REQUIRED FIELDS



To indicate that there is a special reason for not having data in a required field, use:

1. For Numeric Fields

Write NA – This means actively typing NA.

OR

Write a distinct value that would NEVER be used to indicate a real value in the business case.

This option is NOT preferred and only to be used in the case of numeric data types or older computer systems that do not accommodate non-numeric values. Also, the selected value should be consistent and documented in your QA practices.

E.g., -1 or 9999

2. For Text Fields

Write NULL or NA for all text fields, except when storing name of persons.

OR

Write "NOFIRSTNAME" or "NOLASTNAME" when storing name of persons.

This option is preferred for name of persons because Null and Na could be a person's name.

E.g.,
NOFIRSTNAME: missing FIRST name – usually required in the U.S.
NOLASTNAME: missing LAST name – usually required in the U.S.
Leave it blank: missing middle name – not required in the U.S.

Supplement: Metadata Standards

COLUMN AND FILE NAMES

1. Columns and files must be named in a descriptive way – *the names need to make sense to anyone reading the document, including your future self.*

E.g., a column for Patient Disposition should NOT be named "PatDisp." An example of an appropriate name is "PatientDisposition."

2. Avoid space and other special characters, except for underscore – *that is, only use English alphabet characters (A-Z), numbers (0-9), and underscore to avoid confusion and potential query issues.*

E.g., Correct: "patient_disposition" / Incorrect: "patient disposition?"

CALCULATED COLUMNS

1. Calculate columns should be flagged as calculated.

E.g., a column that converts the temperature values of another column from Celsius to Fahrenheit could be named in a way that indicates the temperature was not originally measured in Fahrenheit. An example of a name for this column is "converted_fahrenheit."

2. There should be documentation for how calculated columns were created.

E.g., Preserving Excel formulas instead of replacing them for their resulting values.

Preserving Formula

Inside the cell: =(A2*1.8)+32
Displayed in the cell: 69.8

Coping and Pasting Value

Inside the cell: 69.8
Displayed: 69.8

If you must replace formulas for their resulting values, you can duplicate the tab and save the original formula.

3. Outdated and old versions of calculated columns should be removed from the document.

Appendix C: Project Charter Template

Project Name: *Insert Project Name*

1.0: Project Team	
Project Lead	
Project Team	
Steering Team	*If applicable*
Project Sponsor(s)	
Project Champion	*If applicable*
Equity Analysis Team	*If applicable*

2.0 Problem Statement & Background
What is the problem? How do we know it's a problem? What do we need to know about the issue? What is the context?

3.0 Impact Statement
What will be the impact of addressing the problem outlined above? To the agency, organization, community, etc.

4.0 Scope of Work
What concrete steps do we intend to take to address the problem? Which aspects of the problem will we address, and how?

5.0 Project Resources
What do we need to make this project successful? Data sources, people, review, feedback, etc.

6.0 Final Product
A report, a presentation, a dashboard, etc. What will it consist of? Who is the audience?

7.0 Roles & Responsibilities			
Role	Description	Person	Estimated Time

8.0 Timeline		
Deliverable	People	Due Date

Appendix D: Bibliography

General

"5 Best Practices For Data Cleaning: Increase Your Database". 2022. *Synthio*. Accessed February 8. https://web.archive.org/web/20220208142824/https://synthio.com/b2b-blog/5-best-practices-for-data-cleaning/.

Alley, Garrett. 2018. "What Are Data Silos?". *Alooma.Com*. https://web.archive.org/web/20220208134632/https://www.alooma.com/blog/what-are-data-silos.

Apolitical. 2019. *How To Think Like A Data Journalist*. Video. https://web.archive.org/web/20220208145638/https://vimeo.com/314774736.

Baker, Lee. 2020. "What Is Nominal Data? Definition, Examples, Analysis & Statistics". *Chi-Squared Innovations*. https://web.archive.org/web/20220302220226/https:/www.chi2innovations.com/blog/discover-data-blog-series/nominal-data/.

Benediktas, Benediktas. 2021. "Discrete Vs. Continuous Data: What'S The Difference?". *The Drum*. https://web.archive.org/web/20220225155229/https:/www.thedrum.com/profile/whatagraph/news/discrete-vs-continuous-data-whats-the-difference.

Bhandari, Pritha. 2020. "Data Collection | A Step-By-Step Guide To Data Collection". *Scribbr*. https://web.archive.org/web/20220302173558/https:/www.scribbr.com/methodology/data-collection/.

Bhatia, Manu. 2018. "Your Guide To Qualitative And Quantitative Data Analysis Methods". *Atlan - Humans Of Data*. https://web.archive.org/web/20220308200401/https://humansofdata.atlan.com/2018/09/qualitative-quantitative-data-analysis-methods/.

"Collecting Data". 2022. *Cyfar.Org*. Accessed February 8. https://web.archive.org/web/20220208160923/https://cyfar.org/collecting-data.

"Data Analysis". 2022. *En.Wikipedia.Org*. Accessed March 8. https://web.archive.org/web/20220308171404/https://en.wikipedia.org/wiki/Data_analysis.

"Data Cleansing". 2022. *En.Wikipedia.Org*. Accessed March 7. https://web.archive.org/web/20220307133844/https://en.wikipedia.org/wiki/Data_cleansing.

"Data Ethics". 2022. *Cognizant*. Accessed March 14. https://web.archive.org/web/20220314201820/https://www.cognizant.com/us/en/glossary/data-ethics.

"Data Quality". 2022. *Heavy.Ai*. Accessed March 3. https://web.archive.org/web/20220303201731/https://www.heavy.ai/technical-glossary/data-quality.

"Data Ownership". 2022. *Ori.Hhs.Gov*. Accessed February 8. https://web.archive.org/web/20220208143958/https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dotopic.html.

"Data Viz Project | Collection Of Data Visualizations To Get Inspired And Finding The Right Type.". 2022. *Data Viz Project*. Accessed February 8. https://datavizproject.com/.

Hubbard, Steven. 2020. "Seven Tips For How Public Servants Can Create Better Data Visualisations". *Apolitical*. https://apolitical.co/solution-articles/en/seven-tips-for-how-public-servants-can-create-better-data-visualizations.

"Machine Readable". 2022. Open Data Handbook. Accessed March 3. https://web.archive.org/web/20220303204609/https://opendatahandbook.org/glossary/en/terms/machine-readable/.

"Manage Quality". 2022. *Usgs.Gov*. Accessed February 23. https://web.archive.org/web/20220223142807/https://www.usgs.gov/data-management/manage-quality.

"Metadata". 2022. *En.Wikipedia.Org*. Accessed February 8. https://web.archive.org/web/20220208135234/https://en.wikipedia.org/wiki/Metadata.

"Metadata And Documentation". 2022. Axiomdatascience.Com. Accessed March 3. https://web.archive.org/web/20220228211038/https:/www.axiomdatascience.com/best-practices/MetadataandDocumentation.html.

"Open Data Release Toolkit". 2022. *Datasf*. Accessed February 8. https://web.archive.org/web/20220208150226/https://datasf.org/resources/open-data-release-toolkit/.

Polovets, Leo. 2015. "What Methods Do You Use To Clean Your Data?". *Quora*. https://web.archive.org/web/20220208143638/https://www.quora.com/What-methods-do-you-use-to-clean-your-data.

"Single Source Of Truth". 2022. *Talend.Com*. Accessed April 4. https://web.archive.org/web/20220221190709/https://www.talend.com/resources/single-source-truth/.

Stevenson, Robert. 2022. "Develop A Quality Assurance And Quality Control Plan". *Dataone.Org*. Accessed February 8. https://web.archive.org/web/20220208135552/https://old.dataone.org/best-practices/develop-quality-assurance-and-quality-control-plan.

"Want To Master Your Data? Here's Why You Should Care About Metadata". 2022. Towards Data Science. Accessed March 3. https://web.archive.org/web/20220301205347/https:/towardsdatascience.com/want-to-master-your-data-heres-why-you-should-care-about-metadata-8fcd7754c3b8?gi=baae0cec6995.

"What Is a Project Charter in Project Management?" Wrike. Accessed November 15, 2022. https://web.archive.org/web/20221115212058/https://www.wrike.com/project-management-guide/faq/what-is-a-project-charter-in-project-management/.

"What Is Data Entry?". 2019. *Computerhope.Com*. https://web.archive.org/web/20201027161156/https://www.computerhope.com/jargon/d/dataentr.htm.

"What Is Metadata (With Examples)". 2022. Dataedo.Com. Accessed March 3. https://web.archive.org/web/20220301205154/https:/dataedo.com/kb/data-glossary/what-is-metadata.

"Why Does One Need Clean, Correct And Quality Data?". 2019. Xoriant Cdi. https://web.archive.org/web/20220303201527/https://cdi.xoriant.com/why-does-one-need-clean-correct-and-quality-data/.

Winn, Brad. "Validity, Accuracy, and Reliability." Rogue ABA, July 24, 2019. https://web.archive.org/web/20220311170207/https://www.rogueaba.com/2018/07/12/validity-accuracy-reliability/.

Data Equity

Centering Racial Equity Throughout Data Integration. 2022. PDF. Actionable Intelligence for Social Policy - AISP. Accessed February 2. https://web.archive.org/web/20220120063442/https://aisp.upenn.edu/wp-content/uploads/2020/08/AISP-Toolkit_5.27.20.pdf.

"Data Equity Framework". 2022. *We All Count*. Accessed February 2. https://web.archive.org/web/20220203175710/https:/weallcount.com/the-data-process/.

"Equity Vs. Equality: What's The Difference - Examples & Definitions". 2021. *United Way NCA*. https://web.archive.org/web/20220309190818/https://unitedwaynca.org/blog/equity-vs-equality/.

Gaddy, Marcus, and Kassie Scott. 2020. *Principles for Advancing Equitable Data Practice*. PDF. https://web.archive.org/web/20211006103259/https://www.urban.org/sites/default/files/publication/102346/principles-for-advancing-equitable-data-practice.pdf.

Lee-Ibarra, Joyce. 2020. "Data Equity: What Is It, And Why Does It Matter? — Hawaii Data Collaborative". *Hawaii Data Collaborative*. Accessed February 2. https://web.archive.org/web/20220210194416/https://www.hawaiidata.org/news/2020/7/1/data-equity-what-is-it-and-why-does-it-matter.

Stone, Deborah A. 2020. *Counting: How We Use Numbers to Decide What Matters*. 1st ed. Liveright Publishing Corporation.

Taylor-Powell, Ellen. 1998. *Sampling*. PDF. University of Wisconsin Cooperative Extension. https://drive.google.com/file/d/1akruuiU2_5MAWhKujXFkH09oxJQZRoE7/view.

This page left intentionally blank

Appendix E: Printable Data Management Framework

DATA GUIDE FRAMEWORK *See data guide for additional details.*