Data Quality Guide for Governments

By Stephanie Singer[1]
October 2016

Table of Contents

Introduction

The purpose of this document is to provide a practical guide to any government agency wishing to improve the quality of its data. Whether your data quality project is small (say, improving a single spreadsheet shared by a few users) or large (say, implementing citywide standardization of street addresses across multiple departments and databases), the specific tasks and templates provided here can help you plan, organize and carry out your work.

---

[1] Contact the author at sfsinger@campaignscientific, @sfsinger, 267-414-3119.

We hope that that this guide will be short enough to be approachable, general enough to withstand the test of time in a fast-developing field, and informative enough to prompt useful action.

The Guide would not have been possible without the generous financial support of the Knight Foundation. Chief Data Officer Joy Bonaguro of the City of San Francisco played a key role in the Guide's organization and philosophy. The Guide benefited greatly from Danette McGilvray's *Executing Data Quality Projects*, an excellent book-length reference.

How to Use This Guide

This guide is structured around specific, actionable tasks -- building blocks for a variety of data upgrade projects. Each task has a template spreadsheet designed to guide you through the task and to help you document your work.

The tasks require both technology and business expertise. Someone has to understand the nitty-gritties of the data, and the technology that supports it. Someone has to understand the value of the data in the larger context of the agency's goals and constraints. We strongly recommend working with someone whose strengths complement your own.

The templates are not definitive -- there are many other, probably better, ways to organize process and information. But if you find yourself spinning your wheels on a task, the template should help break the task down into small, approachable units.

The Planning and Buy-In tasks are essential for any project.

Which other tasks you choose from the menu will depend on your particular project. Many data projects fall into one of three categories:

- **Known Opportunity.** Resolving issues with data known to be problematic or combining data from various sources to improve understanding of a particular activity or domain, will often involve:
    - All Planning and Buy-In tasks
    - Several tasks from Preliminaries:
        - Compliance, Security and Backup
        - Life Cycle
        - Stakeholders
    - All Specifications and Integrity Fundamentals tasks
    - Other tasks depending goals and resources
- **New Data**. Planning good data quality practices into a new data collection and maintenance effort, will often involve:
    - All Planning and Buy-In tasks
    - Root Causes from any previous data work

    - Learning from existing systems:
        - Data Decay
        - Inaccuracy Prevention
        - Duplicate Prevention
        - Survey Perceptions
    - Further research to inform design:
        - Assess Opportunities
        - External Standards
        - Consistency Goals
- **Process Improvement**. Planning better data quality practices into existing data collection and maintenance efforts, will often involve:
    - All Planning and Buy-In tasks
    - All Preliminaries, Specifications and Integrity Fundamentals tasks
    - Other tasks depending on goals and resources

List of Tasks

Essential: Planning and Buy-In

The tasks in this section should all be done before touching the data. Unstructured projects with data tend to get bogged down or out of control. Setting deliberate, realistic goals, documenting carefully and communicating strategically will help secure support and maximize the impact of your project.

Task: Goal Statement

Specific, outcome-based goals are powerful. They help a project move ahead efficiently and help garner both internal and external support. What will your jurisdiction, and the people in it, gain from the project? **Template: Goal_Statement**

Task: Project Infrastructure

Basic infrastructure supports efficient use of resources. Documentation allows others (or you, at a future date) to benefit from the work you do now. Reminders help keep the project on track in the face of other claims on your time. **Template: Infrastructure**

Task: Working Group

Relies on Goal Statement. A good working group is key to the success of a data quality project. A small group including expertise in the collection, use and technical aspects of the data can be worth much more than the sum of its parts in all phases: planning, execution and documentation. It is important to include people with deep expertise in the subject of the data. Members of any data coordinating group, data governance group or other mechanism to address data issues across departments will be valuable members of the Working Group. Difficulty assembling a working group (including getting commitments to meet on a regular basis) can be a sign that there is not yet enough support to make the project viable. Consider this task complete after the first meeting of the Working Group. **Template: Working_Group**

Task: Decision-Makers

Relies on Working Group.
Who are the movers and shakers whose support might help the project succeed? Who could hinder the project? Think as broadly as possible. It is particularly helpful to find an executive champion for the project -- someone influential in upper management who can and will promote or defend the project. **Template: Decision_Makers**

Task: Resource Assessment

Relies on Working Group.
A realistic assessment of your resources (with the help of the Working Group) will help you plan projects than can be completed, documented and touted by your department, your agency or the officials who hold the purse strings. **Template: Resource_Assessment**

Task: Effort vs. Payoff

Relies on Working Group.
The Working Group may identify several potential projects. While choosing among these, consider the likely payoffs for each, and the effort and resources that will be required. Think broadly. Money is important, but there are other factors as well -- staffing, public perception, etc.

Don't ignore low-cost, medium-payoff projects, which can be good starter projects, building competence and credibility. **Template:Effort_v_Payoff**

Task: Final Goal Statement

Relies on Working Group and Effort vs. Payoff. After the effort vs. payoff analysis and an assessment of resources available, a new, more focused goal statement may emerge. **Template: Goal_Statement**

Task: Timeline and Deliverables

Relies on Working Group and Final Goal Statement.
Specify the *deliverables* -- the tangible results your project will produce, such as reports, documentation, new dataset, etc. For each deliverable, identify milestones (major drafts, final version, etc.) Even if you're not sure, specify deliverables and timelines. You can always revise them later. Specific deliverables and timelines help keep projects from getting bogged down or out of control. **Template: Timeline_Deliverable**

Task: Tool Selection

Relies on Resource Assessment and Timeline and Deliverables. For anything beyond a basic inventory and data specification you will need software tools. These are available in various forms, including desktop packages that you can run in-house and cloud-based packages. In-house analysts may be able to build reasonable ad-hoc tools. Choosing a tool early will allow time for any necessary procurement and staff training. Find candidate tools by web searches, by asking colleagues for recommendations. "Data Profiling," "Data Cleansing" and "Data Quality" are common keywords used to describe the tools you are looking for. **Template: Tool_Selection**

Task: Communication Plan

Relies on Decision Makers. How will you make sure that key decision-makers understand (at whatever level is appropriate) the goals and needs of your project? Create schedules for keeping each decision-maker up-to-date, and set reminders in your calendar (or whatever system you use to make sure things get done). **Template**: **Decision_Makers**

Task: Data Inventory

Relies on Final Goal Statement and Working Group. The scope of your data inventory -- e.g., within one department, or government-wide on a particular topic -- will depend on the goal of your project. Valuable datasets may be hiding in unexpected places, so the members of the Working Group, or even more widespread brainstorming or surveys may be helpful. **Template: Inventory**

Task: Data Capture

Relies on Data Inventory. Collecting the data and storing it in a form you can work with requires the cooperation of the people who control the data. Planning ahead may prevent unexpected bottlenecks. **Template: Capture_Plan**

Task: Root Causes

One valuable by-product of any data project is insight into root causes of various problems. Even if addressing a particular root cause is outside the scope of the current project, the insight is worth documenting for the benefit of future data projects. **Template: Root_Causes**

Know Your Data

Preliminaries

Task: Existing Documentation

In addition to documentation that may have been created when the current data systems were put into place, there may be documentation from previous data projects. Documentation may save you from duplicating efforts. Documentation may also give insight into root causes of issues. **Template: Documentation**

Task: Compliance, Security and Backup

Best practices in legal compliance, policy compliance, security and backup of stored data are beyond the scope of this guide (and, we hope, beyond the scope of your job). Make sure you know how to contact the people who do this work. **Template: Compliance**

Task: Life Cycle

List all processes and actions that concern the data. McGilvray identifies six life-cycle phases with the acronym POSMAD[2]: Plan, Obtain, Store and Share, Maintain, Apply, Dispose. Make sure to address each of these phases.
Template: Life_Cycle

Task: Stakeholders

Relies on Life Cycle. Who touches the data? Who uses the data? When? How? Why? Reviewing internal processes will help build a comprehensive list of internal stakeholders. Web analytics, data request records and web searches can identify external stakeholders. **Template: Stakeholders**

---

[2] McGilvray, pp. 24-5.

Task: Survey Perceptions

Relies on Stakeholders. What do stakeholders compliment, complain about, wish for? Perception surveys can range from simple web and social-media searches for comments about public datasets to more complicated, statistically sound surveys of internal and external stakeholders. Surveys can elicit general comments or focus on specific aspects such as ease of use, presentation quality or availability.
Template: Perception_Survey

Task: Assess Opportunities

Relies on Stakeholders. If you are in a position to design the data life cycle (perhaps as part of a migration to a new technology, or a brand new data collection initiative) you have an opportunity to design for greater impact and better quality. Eliciting comprehensive, useful information from a wide range of stakeholders is both valuable and beyond the scope of this Guide. There are many resources for what is often called “user-centered design”, including workshops, web materials, books and companies that can guide this process.
Template: Opportunities

Specifications

Good data specifications are essential. Until you know the basic facts about what your data represents and how your data is used, you cannot begin to set goals or estimate resources required for any data quality improvement. Do not skip this section!

All staff who work with the dataset need to know:

- How is the data defined?
- Who are all the in-house users of the data? What are their requirements?
- What do laws and policies require of the data?

The tasks in this section lead you through the process of finding and documenting the answers to these basic questions.

Task: Data Specification Scope

The basic building blocks for using and understanding data are:

- **Standards**, the grammar and spelling of the data world.
- **Reference Data**, fixed, standard lists used in the data.
- **Data Models**, the way the data is organized.
- **Business Rules**, the difference between sense and nonsense, the constraints that real world meaning, plus common sense, put on the data.
- **Metadata**, column definitions, sources and other information about the data or its components.

Tasks below can guide work on these five components of data standardization. The current task is to prioritize this work.
Template: Data_Specs

Task: Standards

Relies on Data Capture. Some standards are explicit in documentation; some can be recognized by looking at actual tables, views and reports. Standards can apply to naming conventions for fields or apply to the content of the data. **Template: Standards**

Task: Reference Data

Relies on Data Capture. A list of US states is an example of reference data. If you've entered an address into an online form, you've probably seen a drop-down menu of states. The form feeds information into a dataset, and the drop-down menu helps to keep the dataset clean. Every entry in the "state" field will be a real state, and each state will have just one spelling. US territories (such as Guam) are sometimes included and sometimes omitted. Consistency of reference data within and between datasets makes merging, duplicating and joining data more effective. **Template: Reference_Data**

Task: Data Model

Data models specify the entities described by the data, the fields describing attributes of entities and the relationships between entities. For a single flat table, a data model is overkill. For more complicated collections of data, data models are essential. Data models are usually presented as diagrams. Because data models are so widely used, there are helpful templates and references available in many formats, including books and web resources.[3] **Template: Data_Model**

Task: Metadata

Relies on Data Capture. Metadata gives basic information about the fields and tables.
Template: Metadata

Task: Business Rules

What constraints do common sense, policy and users put on the data? For example, house numbers on a certain street must be within a certain range; real estate sale prices shouldn't be negative, and no one alive today was born before 1900. Restate business rules in the language of your database (e.g., DOB >= 01/01/1900). Enforcing business rules, usually at the data collection point or during maintenance, will give early warning of both mistakes (such as incorrect data input) or systemic issues (such as the need for a mechanism for government-wide updating of street information when a new block is developed). **Template: Business_Rules**

[3] E.g., https://en.wikipedia.org/wiki/Entity%E2%80%93relationship_model

Integrity Fundamentals

With data specifications in hand, you know what your data is supposed to contain. What does it actually contain? The essential next step is to run some simple tests to look at the basic quality of the content. Discrepancies between the ideal and the actual are signposts toward useful data quality improvement tasks.

Task: Data Profile

Relies on Metadata and Tool Selection. Profiling is done table by table. Take a close look at summary statistics for each field in the table. Patterns that deviate from common sense expectations should be noted for follow-up investigation.
Template: Profile

Task: Field Families

Relies on Metadata, Standards and Tool Selection. Compare fields of the same type. For example, street addresses can take a variety of forms and standards. Use your profiling tool to identify fields that have similar content. Create a comprehensive list of related dataset fields.
Template: Field_Families

Task: Data Decay

Relies on Life Cycle. What lag is caused by your data collection timeline? What events beyond your control compromise your data? How often do they occur? How often is data planned to be collected or revised? What internal agency issues might interfere with execution of the plan? Are the data current and available for use in the timeframe needed by users? In looking for causes of data decay it's useful to consider each stage of the Life Cycle, looking for loss of expertise, process automation, new data uses and changes in the real world that are not captured in the data. Document Root Causes. **Template: Decay_Assessment**

A La Carte Improvements

Deduplication

Task: Deduplication

Relies on Tool Selection, Timeline and Deliverables, Data Model, Internal Standardization and Data Profile. Duplications that waste resources (such as duplicate addresses resulting in wasted postage) come in many forms. Sometimes two rows in a table are strictly identical, or differ only in their primary key. Sometimes a single item in the real world can be described twice in a database with a variety of spellings, or with differences in some fields but not others.

Deduplication of one data table will affect all related tables. Consult your data model. Deleting or merging records (e.g., constituent records) will delete an identifier, which may orphan records in other tables (e.g., contact event records).

Deduplication can be labor-intensive. Deduplication, if done too quickly or carelessly, can result in major damage. Plan, prepare and test methodology before making any irreversible changes to the data. Don't swim alone -- have at least one other knowledgeable person review the methodology and the results of testing the methodology before applying the methodology to the master data. Document any Root Causes found. **Template: Deduplication**

Task: Duplicate Prevention

Relies on Data Profile and Deduplication. Profiling and deduplicating data can uncover patterns of duplication. Whether duplicates were created by people entering single records, or by larger scale data manipulations, it is usually more efficient in the long run to prevent duplications than to fix them. If the duplicate-creating processes are not under your control, the Working Group may be able to help secure the cooperation of the relevant people or departments. **Template: Duplicate_Prevention**

Accuracy & Completeness

How does the data compare to the real world it is supposed to represent? Are the individual records accurate? Are there missing records?

Task: Completeness Assessment

Relies on Tool Selection. Use the data profiling tool to check whether row counts match reality, and to perform other common-sense checks on the data. Any segmentation of your data (by year, by geography, by complaint type, etc.) can be used to look for missing data. **Template: Completeness_Assessment**

Task: Accuracy Assessment

Relies on Resource Assessment. Assessment methods vary widely in cost and effectiveness. The ideal comparison of the data to the real world entities tends to be expensive, while the least expensive options (comparison to proxy sources of data, sampling) may yield misleading results. To picking the right assessment method(s), it helps to know how the results of the assessment will be used. Creating a template for the final accuracy assessment report may help guide the many choices and keep procedures focused on the final, useful result. Document any Root Causes found. Do not update database during the assessment process.
Template: Accuracy_Assessment

Task: Accuracy Correction

Relies on Accuracy Assessment. Update your database with corrections found during the accuracy assessment. **Template: Accuracy_Correction**.

Task: Inaccuracy Prevention

Relies on Accuracy Assessment, Root Causes and Existing Documentation. Any pattern of inaccuracy presents an opportunity to improve the processes that created the inaccuracies. Root Cause analyses from the current or any previous data project may yield other actionable insights. **Template: Inaccuracy_Prevention**

Interoperability

Consistency between the datasets within one department or agency is part of Internal Standardization. Consistency between datasets controlled by a variety of organizations can be trickier to achieve. There is often tension between the value of wide agreement -- about naming conventions, formats, structure of documentation, etc. -- and the value of established culture and practice within departments. The more easily your data can be combined with other data the more value can be extracted, so greater consistency yields public benefits.

Task: Internal Standardization

Relies on Tool Selection, Root_Causes, Field Families, Reference Data and Stakeholders. For each family of fields, pick a uniform standard to be shared by all the fields, specify the reference data (if appropriate) and list the stakeholders who will be affected by the change. As appropriate, obtain the necessary buy-in and then use your data cleaning tool to implement the new standard. Make note of any root causes of obstacles to standardization. **Template: Field_Families**

Task: External Standards

Relies on Field Families, Reference Data and Stakeholders. There are relationships between your data and data held by other organizations, including other governments and government agencies. There are often explicit (or implicit) standards for certain kinds of data. Some related datasets or standards may be well known to you. Others may be known to stakeholders. There may be standards promoted by professional associations, academic institutions, the federal government or consortia of smaller governments. Understanding the landscape of standards and practices allows more informed, deliberate choices about your own data practices and standards. **Template: External_Standards**

Task: Consistency Goals

Relies on Field Families, External Standards and Stakeholders. Achieving consistency with outside standards can be disruptive. Changing standard data definitions, formats or processes

will affect a variety of stakeholders. Careful planning and communication can mitigate disruption. **Template: Consistency_Goals**

Proactive Data Hygiene

Task: Fix Root Causes

Relies on Root Causes. In the course of your work to date, you may have discovered root causes of data quality problems. An ounce of prevention is worth a pound of cure! Addressing some or all of these root causes may well save resources down the road. Improvements can range from simple process improvements within your purview to major initiatives that deserve to be viewed as new data quality improvement projects requiring new planning and buy-in efforts. **Template: Root_Causes**

Task: Data Quality Monitoring

A stitch in time saves nine. Regular monitoring of data quality will help you catch and fix issues before they cause major problems. **Template: Quality_Monitoring**

Task: Culture of Data Quality

Long term, sustainable improvement in data quality is more feasible when people throughout your agency or government understand the value of high quality data and the negative impact of poor quality data. This kind of cultural change can't be forced, but it can be encouraged in various ways. **Template: Data_Quality_Culture**

Resources

Books

Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information , book by Danette McGilvray, Elsevier, Amsterdam, 2008.

Start-Up City, book by Gabe Klein, Island Press, Washington DC, 2015.

- Planning and Buy-In (Introduction)
- Goal Statements (Lesson #2)
- Communication Plan, Working Group (Lesson #3)
- Resource Assessment (Lesson #5)
- Stakeholders (Lesson #6)

Videos

Why Data Quality Is Important (30-minute video), Massachusetts Department of Elementary and Secondary Education, 2014. Accessed 9/20/2016.

- Culture of Data Quality

Implementing Data Quality (Video), DataSourceTV, 2014, accessed 9/21/2016.

- Deduplication (minute 9ff, minute 31ff)
- Data Profiling (minute 19ff)
- Data Decay (minute 20ff)
- Business Rules/Reference Data (minute 22ff)
- Data Quality Monitoring (minute 32ff)

Online Written Resources

Granite Falls Website has a variety of templates and materials developed for data quality projects. Accessed 10/11/2016.

Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 1.0), Meredith N. Zozus, et al., 2014. Accessed 9/20/2016.

Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data, Wayne W. Eckerson, TDWI Report Series, 2002. Accessed 9/20/2016.

- Root Causes (pp. 12-3)
- Standardization (pp. 13-4)
- Culture of Data Quality (pp. 14-8 and 25-6)
- Fix Root Causes (pp. 18-9)
- Data Profile (pp. 19-22)
- Monitor Data Quality (pp. 24-26)

Wikipedia Articles

- Data Model (Relationship Model)
- Assess Opportunities (User-Centered Design)

Data Quality and Data Cleaning: An Overview (Lecture Slides), Theodore Johnson, 2004. Accessed 9/20/2016.

- Root Causes (slides 9-25)
- Monitor Data Quality (slides 36-41)
- Data Profile (slides 35-72, 87-94)
- Deduplication (slides 82-5)
- Metadata (slides 95-7)