Data Quality

Child Welfare
Workforce Analytics Institute

Data-driven decisions cannot be effective if they are based on poor-quality data. Thus, it is important to ensure that any data that will be used for workforce analytics are accurate and can be trusted.

Data Quality Criteria

There are many potential data quality criteria to consider. Below are 9 selected criteria that are relevant for workforce data. If there are others that are also important to you, they should be included as well. These questions should be asked of every variable for which data are collected.

1. *Relevance:* Are the data useful and relevant for business needs?
2. *Granularity:* Are the data at a sufficient level of detail for answering important questions and making effective decisions?
3. *Completeness:* Are there values entered for every case, or are there missing values?
4. *Validity:* Do the data entered conform to a specific set of rules, such as data type, format, codes, etc.?
5. *Accuracy:* Do the data correctly represent reality?
6. *Timeliness:* Are data entered in a timely fashion or are there significant lags? Are the data updated or do they stay the same over time? If they are updated, are they overwritten or is a history retained?
7. *Uniqueness:* Is this the only record of this information, or is the information duplicated in this or another data source?
8. *Consistency:* If the information is duplicated, are the data consistent, or could there be differences?
9. *Availability:* Are the data accessible to the people who need them?

Assessing Data Quality

There are numerous ways to assess the quality of current data, ranging from simple and cursory to sophisticated and thorough.

1. The first step is to determine whether there are existing data standards, dictionaries, or codebooks that provide information about the structure, contents, and layouts of your data files. These may not guarantee high-quality data, but they provide a benchmark for what the data should look like.
2. Though they won't be able to speak to every data point, people who are familiar with the data should be able to provide some information about the various data quality criteria.
3. The software that contains the data may include a means of readily viewing the characteristics of your data, so check for those options.
4. A visual scan of the data can provide some preliminary information about a few of the criteria. If there is a large amount of data, strategic sampling can ease the burden.
5. A frequency distribution for each variable will show the number of valid and invalid values, missing data, outliers, and the format of existing data.
6. The most systematic and comprehensive means of exploring existing data is to use data profiling tools, which examine data structure, content, and relationships. Excel's *Get and Transform* function will handle

some of the more basic aspects of data profiling. There are also a variety of open source and commercial software, which can perform very sophisticated data profiling tasks. A few resources include:

Excel

Power Query in Excel (formerly Get & Transform)
Power Query Overview
Quick Data Profiling Solution
Find data issues in the Power Query Editor
Data Profiling in Power Query

Open Source and Commercial Data Profiling Tools

Trifacta Wrangler
Data Profiling: A Secret Weapon in Your Fight Against Bad Data
Start Wrangling
Getting Started with Trifacta Wrangler

Google DataPrep
Dataprep by Trifecta
Dataprep by Trifecta How-To Guides

Lists of other tools
What Is Data Profiling? Process, Best Practices and Tools
5 Best Data Profiling Tools and Software Solutions

As you detect data quality issues, it's a good idea to track details about those issues. For suggestions about what to track, see The ultimate guide to a Data Quality issues log.

Correcting Data

Before initiating steps to clean or correct data, it is important to consider the costs and benefits of doing so. If cleaning existing data will take a significant amount of effort, you may consider just focusing on efforts to improve the quality of future incoming data. If you decide that cleaning will not require too much time and effort or that the benefits will outweigh the costs, the next step is to determine the reason for each data quality issue. This will help with both correcting and preventing poor-quality data. After there is a solid understanding of the reasons for the quality issues, then you can use tools and processes for correcting the data.

Understanding the Causes of Poor-Quality Data

Examination of data in the previous step can reveal or hint at some causes, but it will likely be insufficient. It is important to also connect with users and data managers to better understand potential underlying issues. Possible causes include (but are not limited to) the following:

- No (or poor) processes for obtaining data that need to be entered
- Expectations for data entry (who, what, when, where, how) are absent, inconsistent, or unclear
- Data are not seen as important, or priorities are unclear
- Too much data to enter, or too many steps to get them ready or configured for entry
- Architecture of data system is too permissive or too restrictive (e.g., inaccurate values allowed, fields are too long or not long enough, choices don't always fit the situation, a field isn't available so another one is used instead, a field is mandatory even though it might not be applicable)
- Need to enter data more than once, in more than one place, potentially by more than one person

- Human errors
- Poor user interface
- Data decay, no triggers or processes for updating
- Data extraction errors
- Data purging
- System errors
- System upgrades

Again, as you discover the causes of data quality issues, it's a good idea to track that information in a log.

Tools and Processes for Correcting Data

- Data can be corrected through the use of a spreadsheet application like Excel or through software that is specifically dedicated to improving data quality.
- Many of the data profiling tools listed in the previous section can also be used for cleaning data, and there are additional commercial data cleansing tools available.
- Some of the more common functions of data cleansing tools include (McGilvray, 2008):
  - Standardizing: transforming data into standard formats
  - Parsing: separating text or string data into parts (e.g., separating first and last name into two separate fields)
  - Enhancing: updating, correcting, or adding new information to existing information
  - Matching: linking associated records
  - De-duplication: identifying duplicate records
  - Merging: combining duplicate records
- Note that even if data are accurate and "clean," they may not be in the proper format for analysis (e.g., analyses on text or string data are very limited), so data cleansing tools can be useful even for otherwise high-quality data.
- Automated data cleaning processes may need to be supplemented or followed by manual efforts, particularly for null values (i.e., fields for which data exist but have not been entered).
- Keep in mind that data corrections should be done with caution. Depending on the complexity of the data and the scope of the corrections, the process may introduce errors if not executed carefully.
- Again, as you transform any data, it's a good idea to track that information in a log.

Improving Future Quality Data

After, or in lieu of, correcting poor-quality data, it is important to implement strategies to improve the quality of future incoming data. The data quality criteria on page one serve as an important guide, and there are a number of suggested strategies to support the process of achieving them.

Business Rules

- Create data quality rules that indicate what constitutes valid data (e.g., data type, format, values)
- Create documents that outline the business rules for data, such as data definitions guides and data dictionaries, so everyone has a shared understanding of the rules

Architecture

- Apply the business rules to the design of your data system, so data quality is built into the system (e.g., a field only allows numbers and not text, null values not permitted)

- Create data validation rules in your data system (i.e., the system rejects entries that don't fit the rules)
- Create duplicate detection rules
- Keep data in one system, to avoid double accounting, which duplicates efforts and records and will likely lead to inconsistencies

Processes

- Control incoming data
- Profile data as a first step before an analysis or report
- Regular monitoring, possibly through data quality KPIs or dashboards

People

- Determine who has data ownership and access; find the right balance that ensures necessary limits but doesn't create siloes and force people to resort to redundant processes and data systems
- Establish clear roles regarding who will do what with which data
- Ensure that everyone has the necessary knowledge, skills, and resources to fulfill their data-related roles (e.g., training, accessible manuals, FAQs)

Resources/Additional Reading

McGilvray, D. (2008). Executing data quality projects: Ten steps to quality data and trusted information. Morgan Kaufmann.

Funded through the Department of Health and Human Services, Administration for Children and Families, Children's Bureau, Grant #HHS-2016- ACF-ACYF-CT-1178. The content of this publication does not necessarily reflect the view or policies of the funder, nor does mention of trade names, commercial products or organizations imply endorsement by the US Department of Health and Human Services.