Describing and Disseminating the FCSM Data Quality Framework

Darius Singpurwalla
National Center for Science and Engineering Statistics
Federal Committee on Statistical Methodology Member

Friday, December 1st, 2023

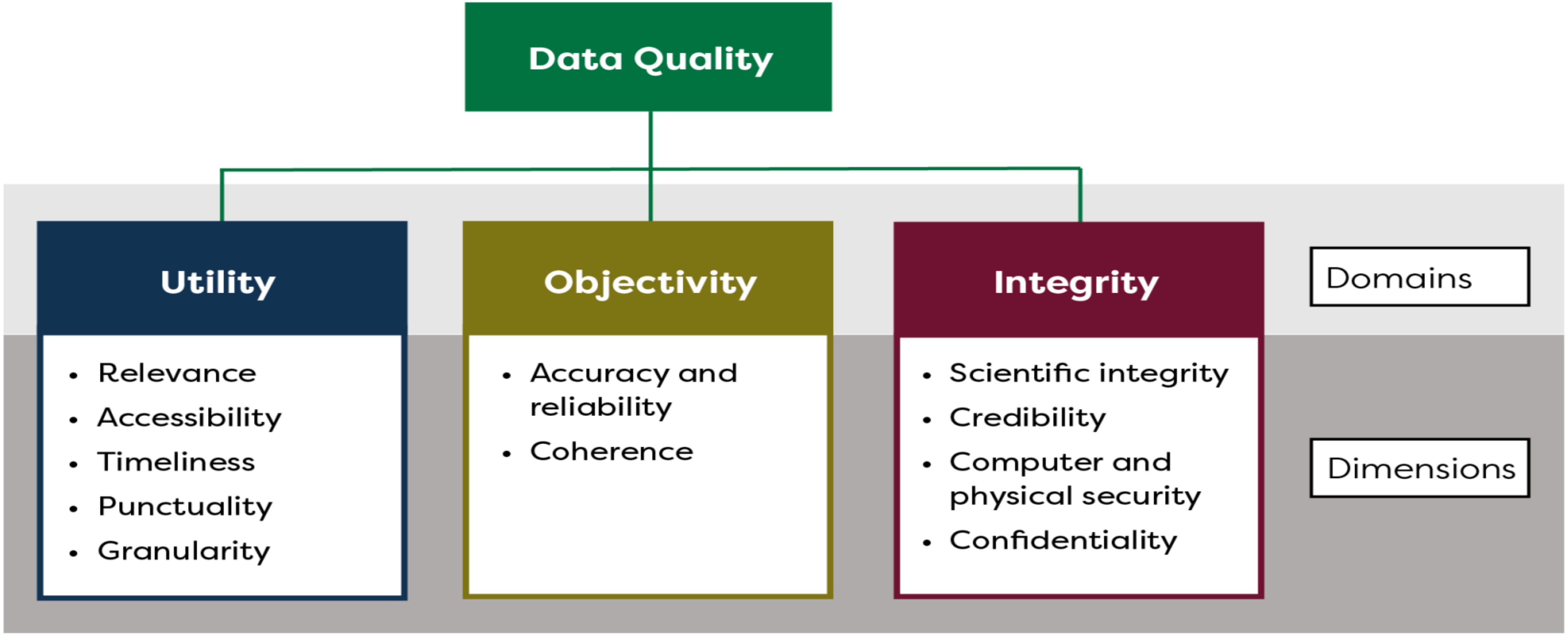
Agenda

Goal of the presentation is to introduce and promote the FCSM Data Quality Framework.

1) **Introduction to the FCSM Data Quality Framework**
    1) Review of Domains and Dimensions of the Framework
2) **Applications of the Data Quality Framework**
    1) Case Study Manual
    2) Scorecard
3) **Further Dissemination of the Data Quality Framework**
    1) Working with the scorecard
    2) Outreach to other agencies

Overview of the FCSM Data Quality Framework

Figure ES 1. The FCSM Data Quality Framework



Data Quality

Utility

- Relevance
- Accessibility
- Timeliness
- Punctuality
- Granularity

Objectivity

- Accuracy and reliability
- Coherence

Integrity

- Scientific integrity
- Credibility
- Computer and physical security
- Confidentiality

Domains

Dimensions

Connection Exercise

- Think about the potential impact that poor data quality can have when….
    - A company relies on customer data to target its marketing campaigns. Due to poor data quality, the marketing team sends promotional materials to outdated or incorrect addresses, leading to a waste of resources and potentially irritating customers.
    - A financial institution uses inaccurate data for risk assessments and investment decisions. Flawed data may include incorrect credit scores, outdated financial histories, or miscalculated risk indicators.
    - A customer contacts a company's support center for assistance, but due to poor data quality, the customer service representative lacks accurate information about the customer's previous interactions and issues.

FCSM: Case Study Handbook

- Consists of seven case studies
- Follows this format:
    - Description of the data being assessed for quality.
    - How the FCSM data quality framework compares with previous efforts to assess data quality
    - Description of implementation including human capital, technology needed, and cost
    - Assessment of the source using the framework
        - Domains/Dimensions
    - Lessons Learned / Sustainability

Overview of the Data Quality Case Studies

Darius

NCSH Linked Mortality File

Overview: This case study uses the Framework for Data Quality to assess the quality of National Center for Health Statistics' (NCHS) Linked Mortality Files (LMFs), which blends survey and mortality data.

Since **blended** data can increase the **disclosure** risk of a dataset, this case study describes the procedures and methods used by NCHS to systematically address the dimension of confidentiality ,in addition to other dimensions of the framework.

CPI – Crowdsourcing Gasoline Prices

Overview: This case study describes the data quality assessment of a new method for collecting Consumer Price Index (CPI) gasoline price data from retailers or data aggregators instead of a sample survey at BLS. While crowdsourcing data directly from retailers leads to efficiencies in both collection efforts and costs, the method also introduces the potential for increased errors in collection. This case study highlights how the BLS mitigates threats to the **accuracy** and **reliability** of these crowdsourced data and is using the framework to guide the expansion of alternative data into CPI estimation.

Overview of the Data Quality Case Studies

Darius

Motor Carrier Inspection Data

Overview: This case study evaluates data collected from the U.S. Department of Transportation (DOT)'s new roadside inspection tool, SafeSpect, highlighting the **utility** domain in the DQ framework and, more specifically, the **relevance** and **timeliness** dimensions. With respect to relevance, the tool continues to provide highly relevant data that support the Federal Motor Carrier Safety Administration's (FMCSA) mission. The case study describes in some detail how the new tool improves the timeliness of making the data available for review.

Physical Activity Monitoring from NHANES

Overview: This case study assesses the **utility** of physical activity monitor (PAM) data collected in the National Health and Nutrition Examination Survey (NHANES), which is conducted by the National Center for Health Statistics (NCHS) within the Centers for Disease Control and Prevention (CDC). This study focuses specifically on evaluating the **accessibility** of large data files that require subject matter expertise to analyze.

Overview of the Data Quality Case Studies

Darius

Program Evaluation Data

Overview: Federal agencies interested in assessing the effectiveness of government programs and services often face limited resources. Thus, agencies often rely on administrative data that are already collected by the government for a different purpose and use the data to aggregate program outcomes and estimate policy impacts. Evaluators have become entrepreneurial in identifying data sets that will support answering research questions of interest through rigorous program evaluation, which may necessarily involve matching across multiple sources or require some supplemental data collection. It is imperative that researchers are transparent about the fitness for use of existing data for program evaluation, to strengthen credibility of study design and findings, and to account for other factors including data privacy. This case study examines how the FCSM Framework for Data Quality serves as a tool to determine the utility of administrative data in determining program effectiveness and provides considerations for the data's use that are paramount to upholding rigor and ethics as principles of program evaluation.

Conceptualizing a New Study

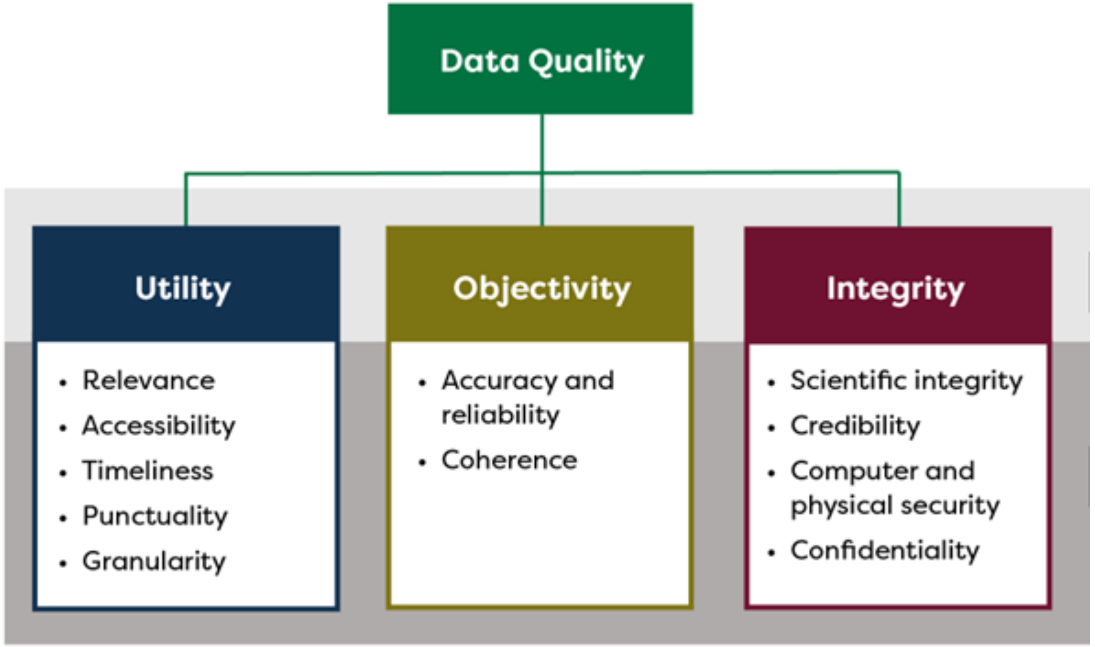
Overview: In this case study, the FCSM Framework for Data Quality was used to anticipate potential data quality threats during the planning stages for a new study that is designed to measure the nation's perceptions of the science and engineering (S&E) enterprise.

Reflections on the Case Studies

- Provide detailed information about the advantages & limitations of a particular data source.
  - Advantages
    - Nuance
    - How threats were addressed
    - Comparisons to other data quality evaluations
  - Limitations
    - Potential selection bias
      - Provides information on the dimensions as chosen by the author
      - ***Not all threats may be addressed in the write-up or the write-up focuses heavily on one set of dimensions (e.g., accuracy and reliability)***
    - Time intensive to generate and can be less generalizable than other types of analysis.

Data Quality Scorecard

Darius



- Relevance
- Accessibility
- Timeliness
- Punctuality
- Granularity

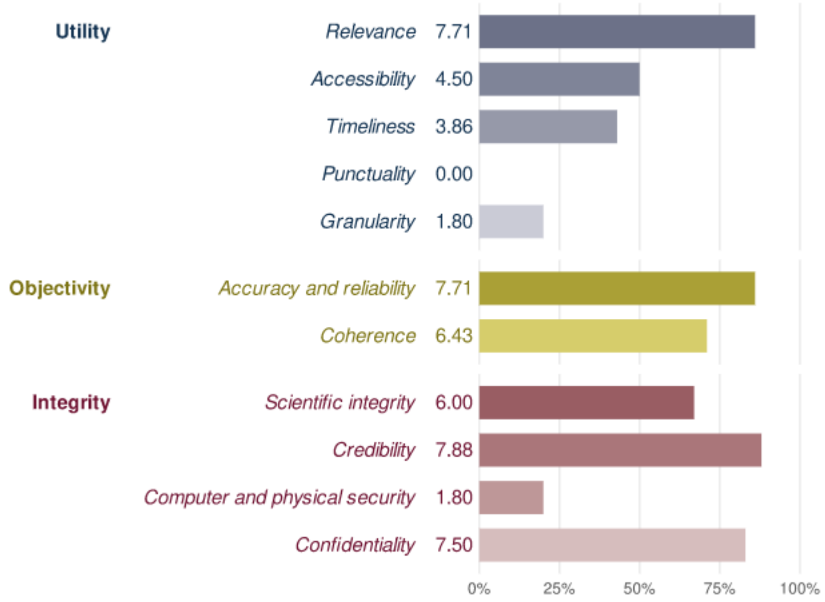
- Accuracy and reliability
- Coherence

- Scientific integrity
- Credibility
- Computer and physical security
- Confidentiality

SCORE SUMMARY

Overall Score: 56/100

Each dimension is scaled to 9 points. One point is given by default.



Reflections on the Scoring Data Quality

- Multiple stakeholders with varying expertise should collaborate to complete the scorecard.
  - Familiarity with data source.
  - Knowledge of subject matter applications
  - Statistical Expertise
- Scores are use-case specific.
- Data documentation may be necessary for stakeholders to complete this scorecard.
- Comparing data quality scores across multiple data sources can illuminate the candidate sources strengths and weaknesses.
- Data producers can use the scores as benchmarks to improve future collections.

Future Initiatives

- Next Steps
  - Further work on scorecard
    - Testing scorecard with the case studies.
    - Comparative scoring (multiple people evaluating same dataset)
    - Scoring as a benchmark
  - Doing outreach with other agencies
    - Goal to get a standardized way to discuss data quality.