A Data Quality Scorecard to Assess a Data Source's Fitness for Use

Elizabeth Mannshardt

FCSM

10/25/2023

NATIONAL CENTER FOR SCIENCE AND ENGINEERING STATISTICS
NATIONAL SCIENCE FOUNDATION

John Finamore; NCSES
Julie Banks, F. Jay Breidt, Zachary H. Seeskin, Kiegan Rice; NORC at the University of Chicago

Overview

Motivation

Assessment

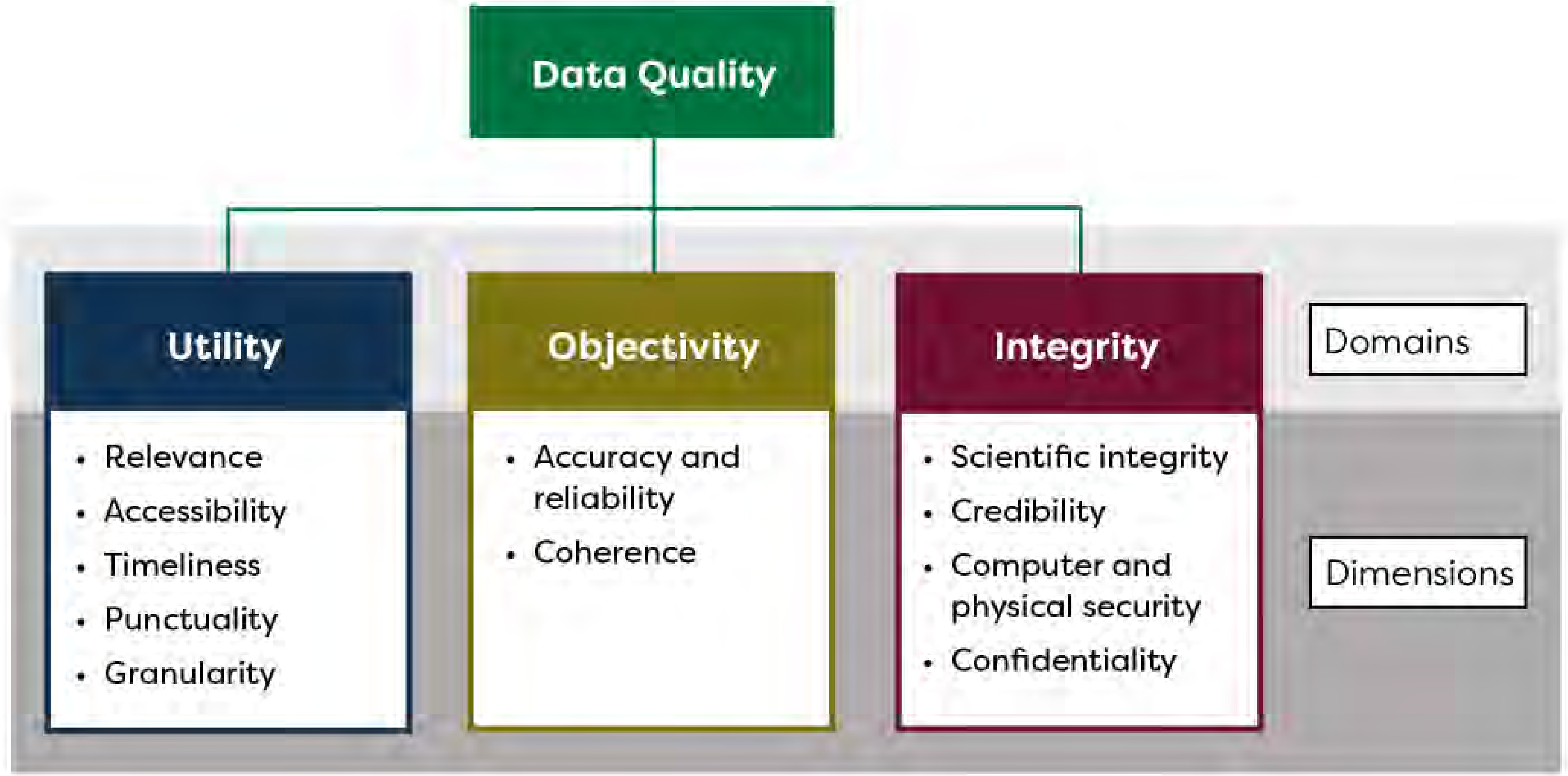
Design Features

Discussion

Motivation

Motivation

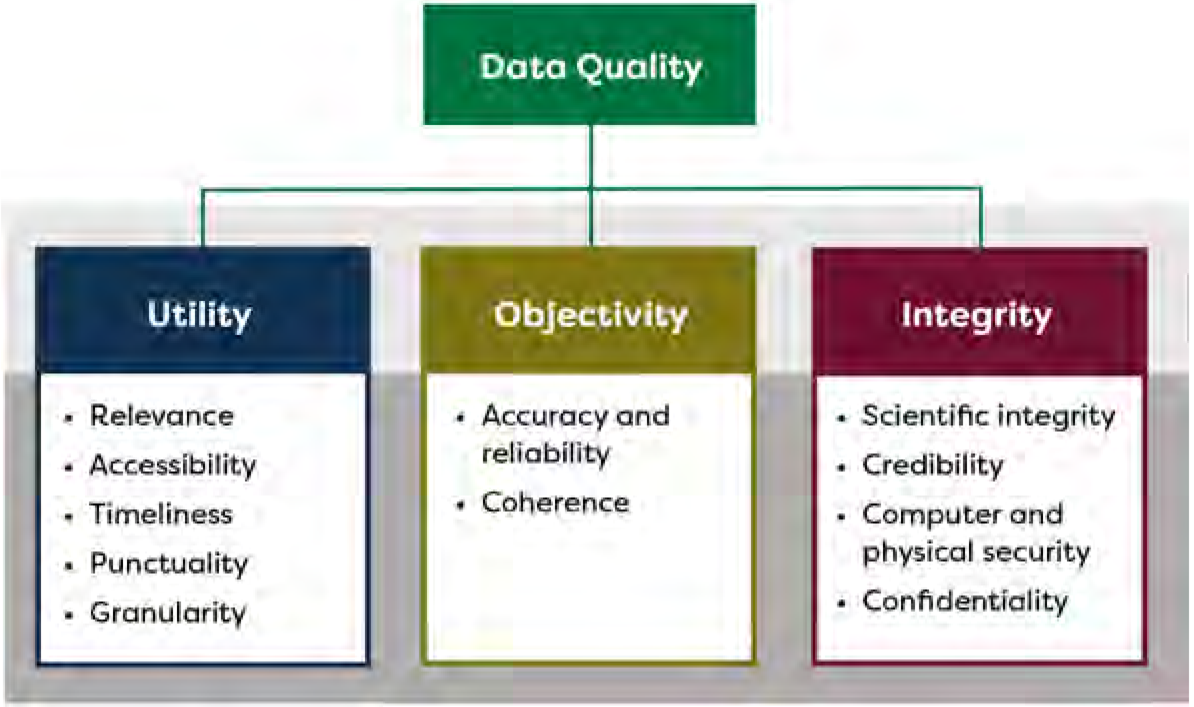
The FCSM Framework for Data Quality provides valuable guidance to assess data quality. Tools are needed to support application of the Framework to data files.



Federal Committee on Statistical Methodology. 2020. *A Framework for Data Quality*. FCSM 20-04.

Motivation

Data quality assessment should be user-friendly, objective, and reproducible.

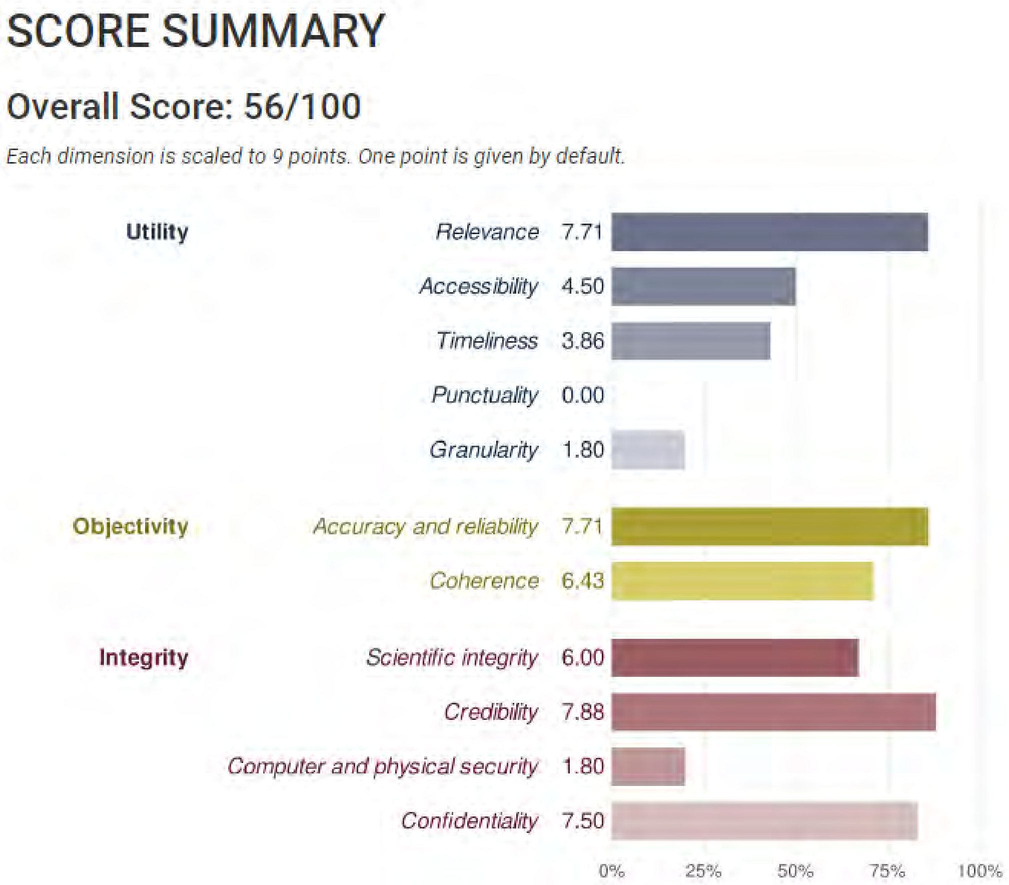


Data Quality

Utility
- Relevance
- Accessibility
- Timeliness
- Punctuality
- Granularity

Objectivity
- Accuracy and reliability
- Coherence

Integrity
- 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.

Motivation – NCSES’s federal clearinghouse role

A Variety of Data Sources

- NCSES official statistics
    - Universities
    - Government agencies
    - Businesses
    - Individuals
- Census, BLS, NCES, etc.
- International, OECD data
- Bibliometric data
- Patent data information

The State of U.S. Science & Engineering

2022

Science & Engineering Indicators

NATIONAL SCIENCE BOARD

Diversity and STEM

Women, Minorities, and Persons with Disabilities

2023

National Center for Science and Engineering Statistics
Directorate for Social, Behavioral and Economic Sciences
National Science Foundation

Motivation

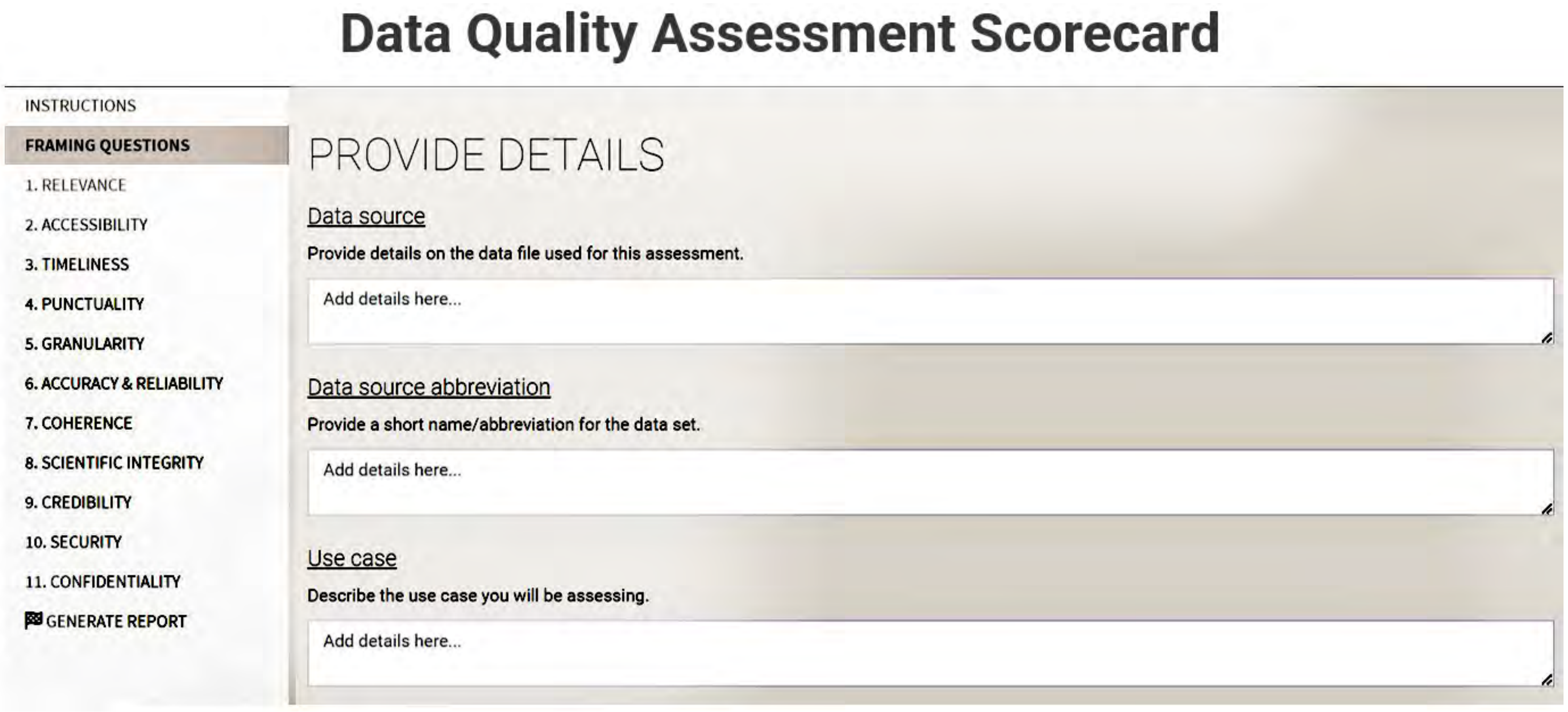
- Quality assessment is critical to evidence-building
- Challenges
  - Complex policy and research questions
  - Additional/evolving data sources and data types
- Needs
  - Consistent terminology
  - A user-friendly tool to assess data capabilities and limitations
  - Method to operationalize the FCSM Data Quality Framework

Assessment

Assessment

Fitness for use and Framing Questions

- Use case
- Data reference period
- Temporal scale
- Assessment reference period



Data Quality Assessment Scorecard

PROVIDE DETAILS

Data source

Provide details on the data file used for this assessment.

Add details here...

Data source abbreviation

Provide a short name/abbreviation for the data set.

Add details here...

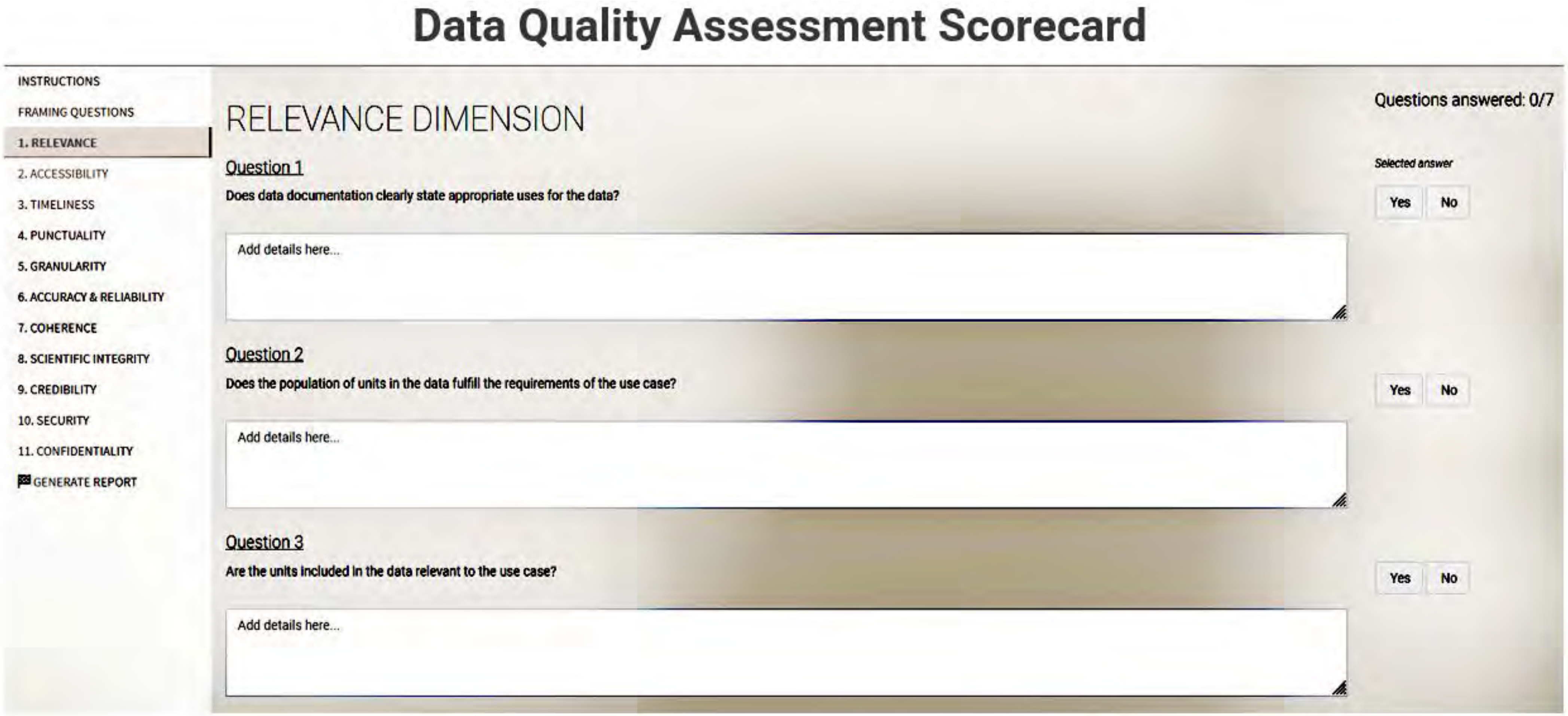
Use case

Describe the use case you will be assessing.

Add details here...

Assessment

Operationalizing the FCSM Data Quality Framework



Data Quality Assessment Scorecard

RELEVANCE DIMENSION

Questions answered: 0/7

Question 1

Does data documentation clearly state appropriate uses for the data?

Selected answer

Add details here...

Question 2

Does the population of units in the data fulfill the requirements of the use case?

Add details here...

Question 3

Are the units included in the data relevant to the use case?

Add details here...

Assessment

Example Scorecard Questions

- **Relevance (Utility Domain):**
  Does data documentation clearly state appropriate uses for the data?
  Are the units included in the data relevant to the use case?

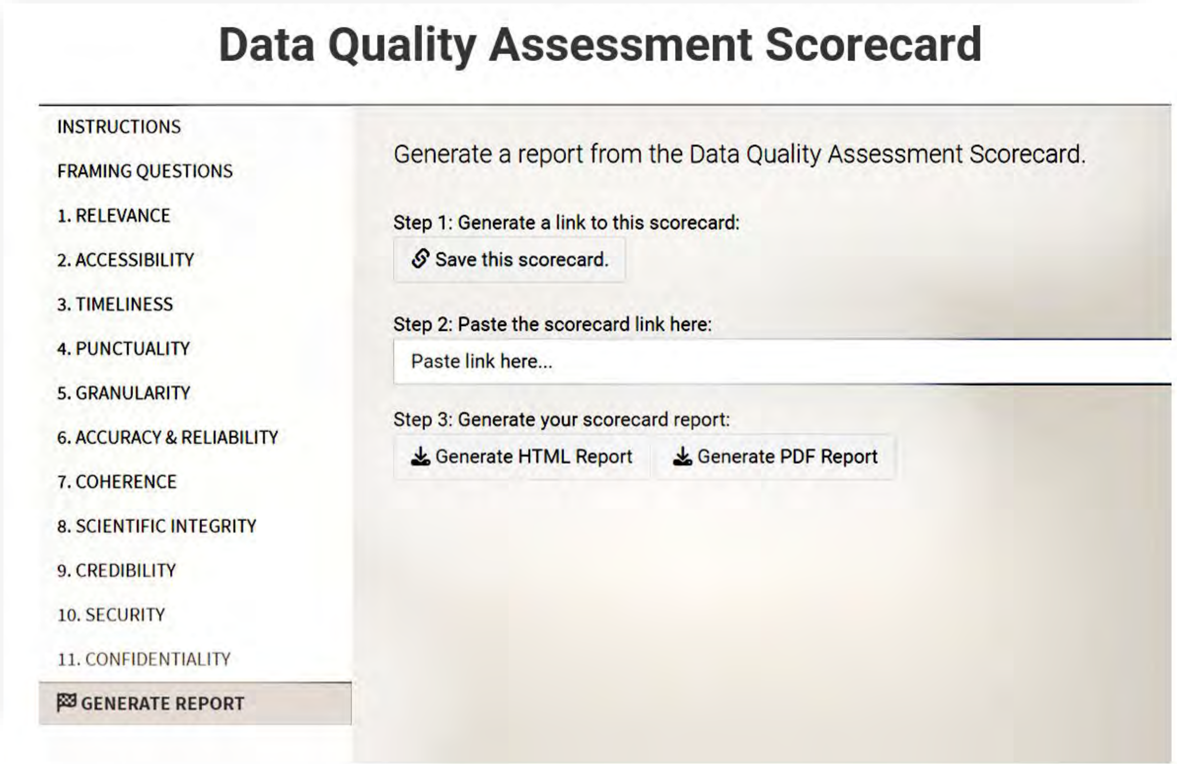
- **Coherence (Objectivity Domain):**
  Are key variables defined in the same way for each record in data, including across subgroups?
  Are key variables measured using same techniques as similar constructs from external data sources?

- **Credibility (Integrity Domain):**
  Is the data producer a non-partisan organization?
  Is there documentation on all of the following topics:
  (a) Data collection methods or data generation processes,
  (b) Methodological assumptions
  (c) Data error or mitigations to reduce error
  (d) Data limitations.

Design Features

Design Features

After completing the evaluation, the users can generate an HTML or PDF report that can be shared as part of the supporting documentation for a project.



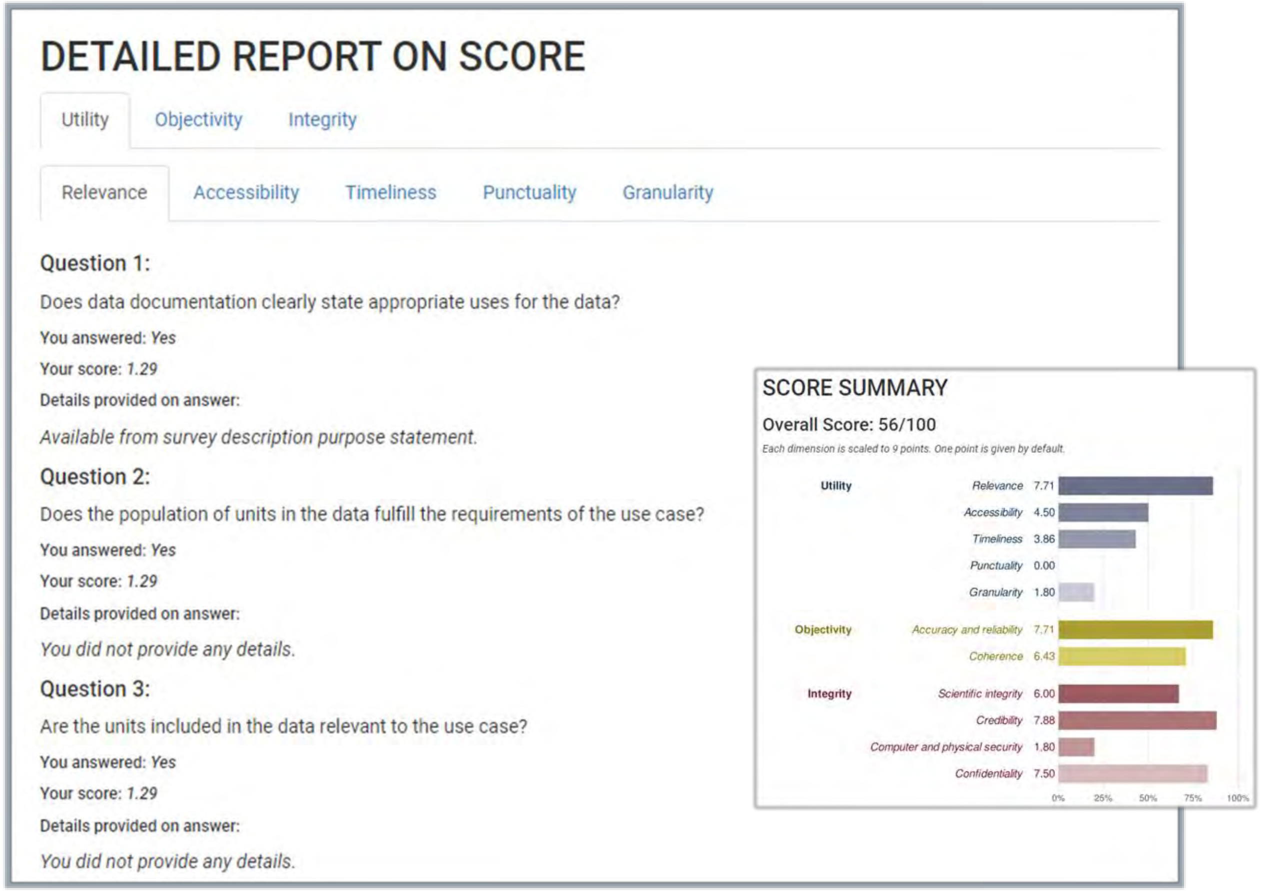
Data Quality Assessment Scorecard

Generate a report from the Data Quality Assessment Scorecard.

Step 1: Generate a link to this scorecard:

Step 2: Paste the scorecard link here:

Step 3: Generate your scorecard report:



DETAILED REPORT ON SCORE

Question 1:

Does data documentation clearly state appropriate uses for the data?

You answered: *Yes*

Your score: *1.29*

Details provided on answer:

Available from survey description purpose statement.

Question 2:

Does the population of units in the data fulfill the requirements of the use case?

You answered: *Yes*

Your score: *1.29*

Details provided on answer:

You did not provide any details.

Question 3:

Are the units included in the data relevant to the use case?

You answered: *Yes*

Your score: *1.29*

Details provided on answer:

You did not provide any details.

SCORE SUMMARY

Overall Score: 56/100

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

Design Features

With the R Shiny and R Markdown setup, the tool readily supports edits

- Dashboard content maintained in Excel metadata spreadsheet
  - Domain and dimension details
  - Question text
  - Question details and information
- Updates to metadata in spreadsheet automatically reflected in dashboard
- User interface elements generated programmatically using functions that pull text from Excel metadata

Discussion

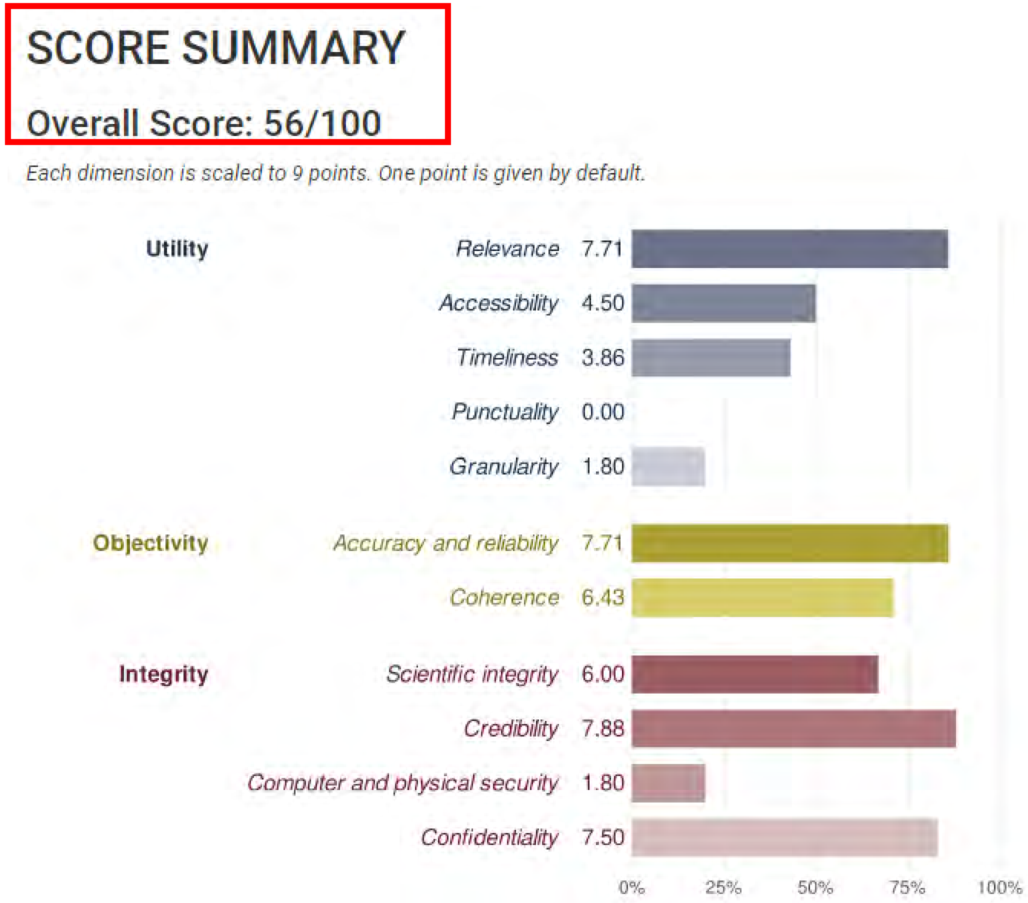
Discussion

We recommend that users with different areas of expertise collaborate to complete each scorecard. Areas of expertise needed include:

1. Familiarity with or investigation of data source
2. Knowledge of subject matter applications
3. Statistical expertise

Discussion

Score is **use-case specific**



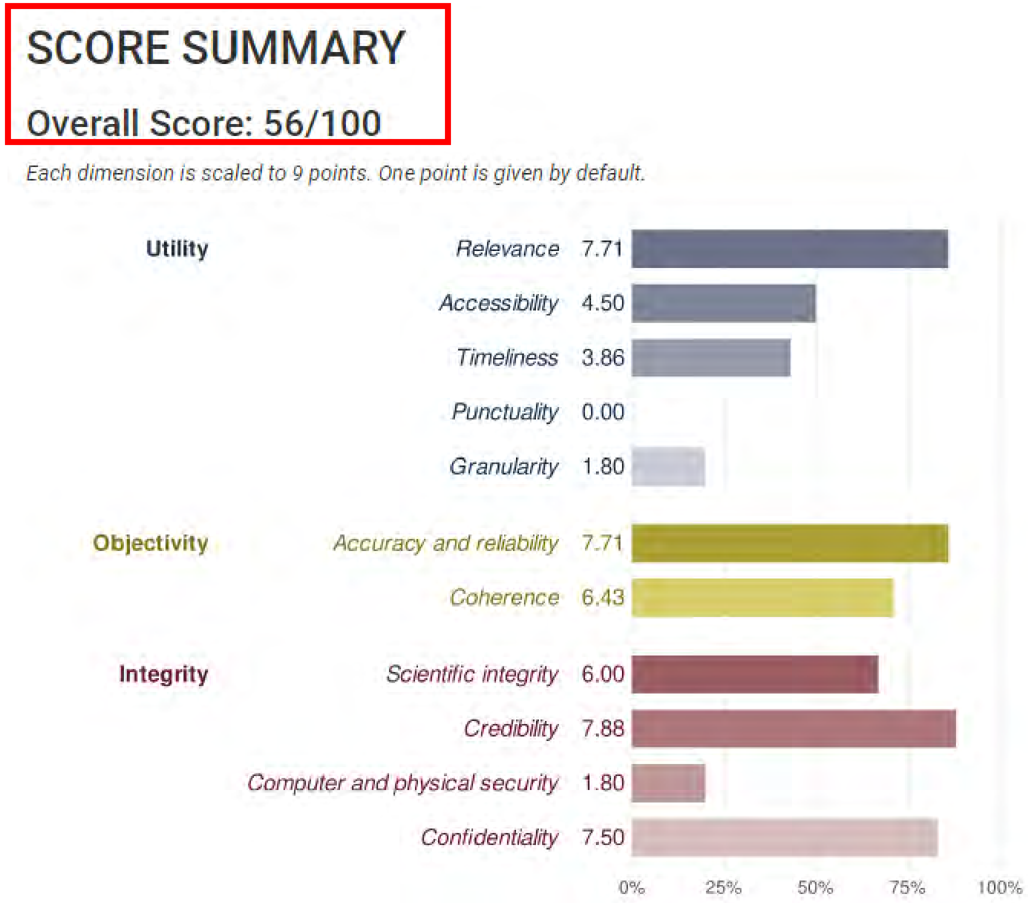
SCORE SUMMARY

Overall Score: 56/100

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

Discussion

Data documentation should inform responses to scoring items and may be more complete for some data sources than others (survey vs non-survey)



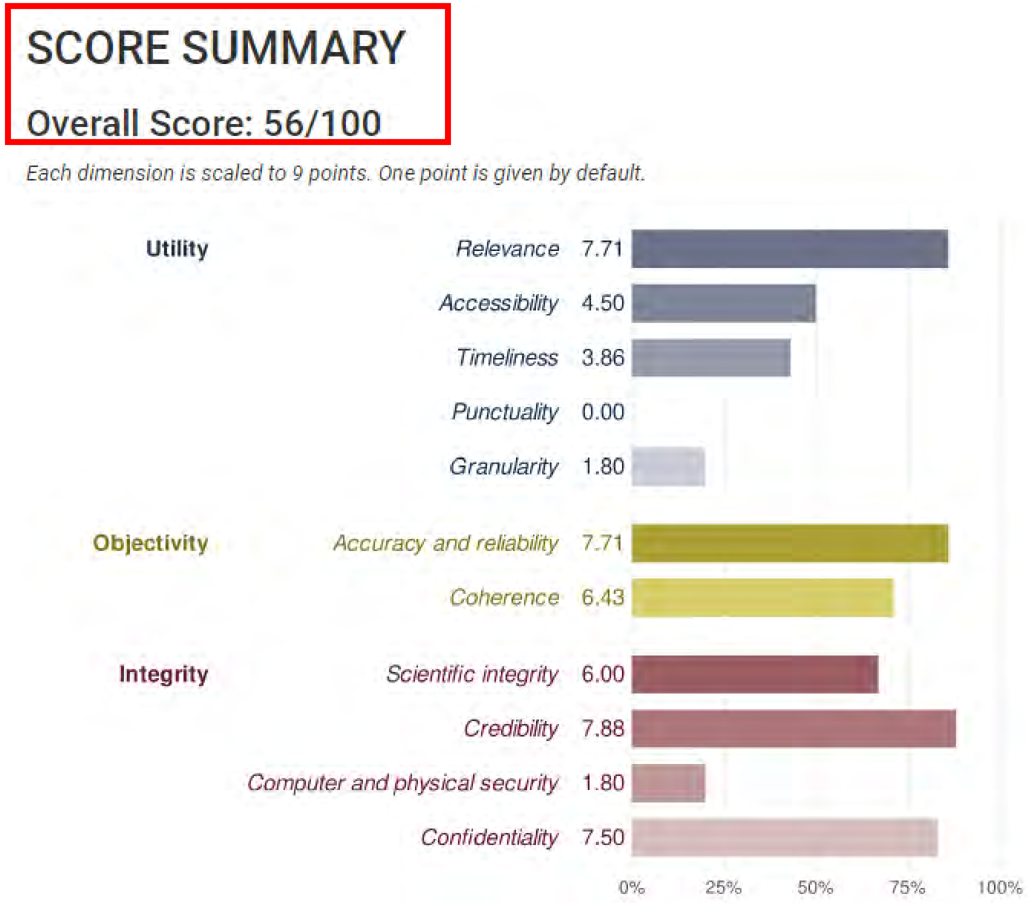
SCORE SUMMARY

Overall Score: 56/100

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

Discussion

Comparing data quality scores can reveal strengths and limitations of different data sources, including between survey and non-survey data.



SCORE SUMMARY

Overall Score: 56/100

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

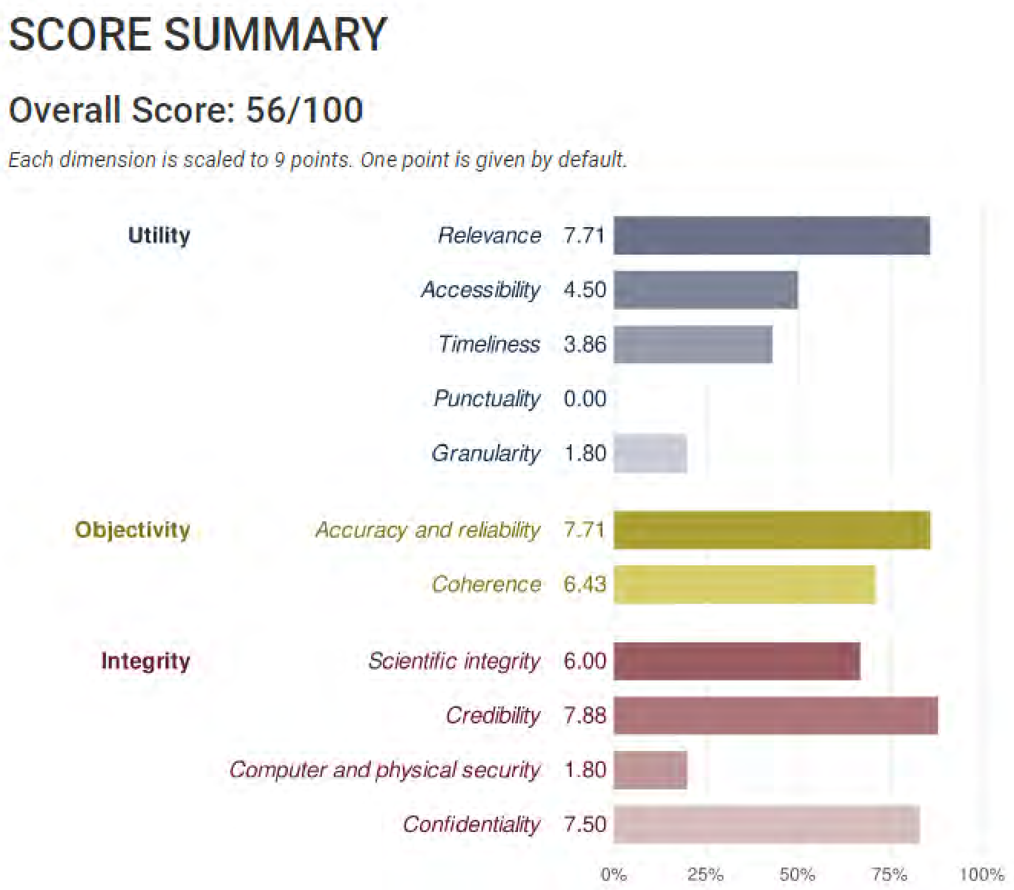
Discussion

Because quality of a data file is based on a specific use case, it can be valuable to evaluate data quality with alternate use cases.

- Use cases may vary in needs for timeliness, granularity, or accuracy
- Aspects of ‘relevance’ dimension may vary among use cases
- The scorecard can load a prior use case scorecard for the same data source as a starting point

Discussion

Instances where non-survey data sources have advantages



SCORE SUMMARY

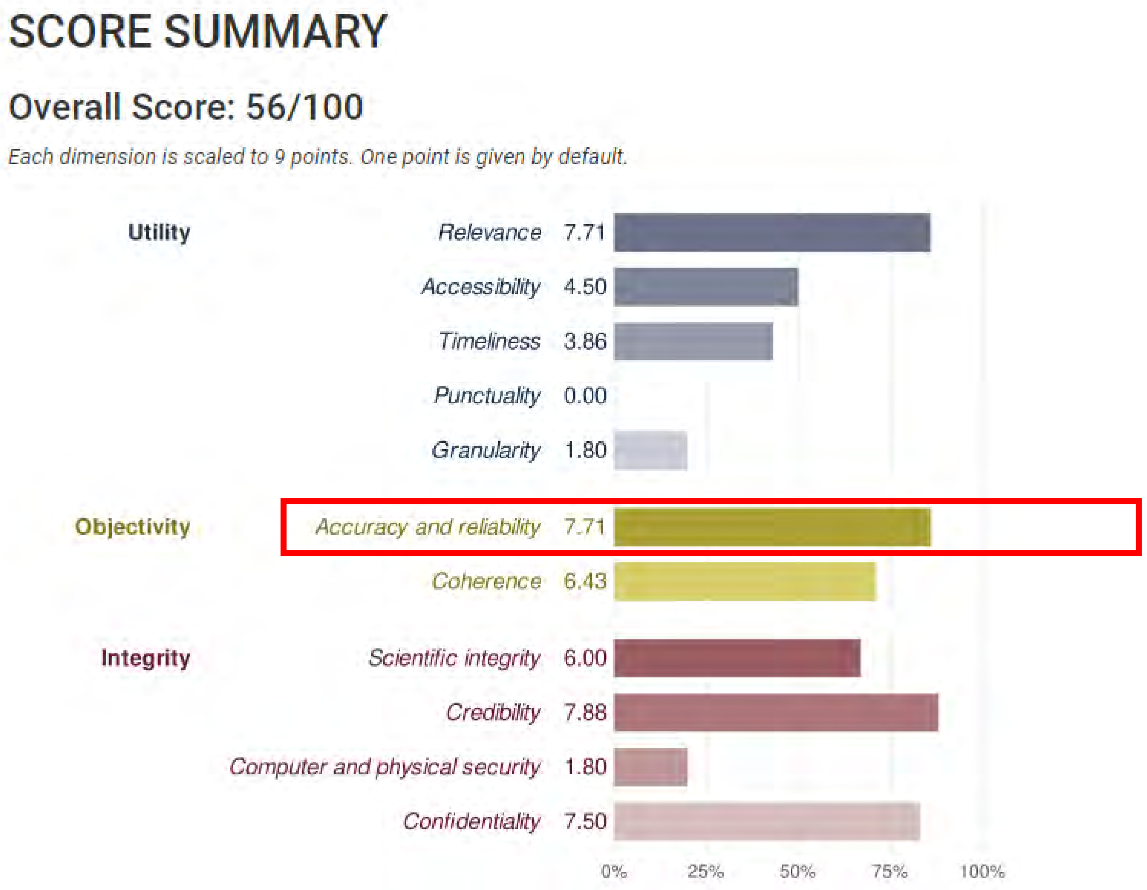
Overall Score: 56/100

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

Utility	*Relevance*	7.71
	Accessibility	4.50
	Timeliness	3.86
	Punctuality	0.00
	Granularity	1.80
Objectivity	*Accuracy and reliability*	7.71
	Coherence	6.43
Integrity	*Scientific integrity*	6.00
	Credibility	7.88
	Computer and physical security	1.80
	Confidentiality	7.50

Discussion

Consider use case: Some data sources have advantages with **accuracy**



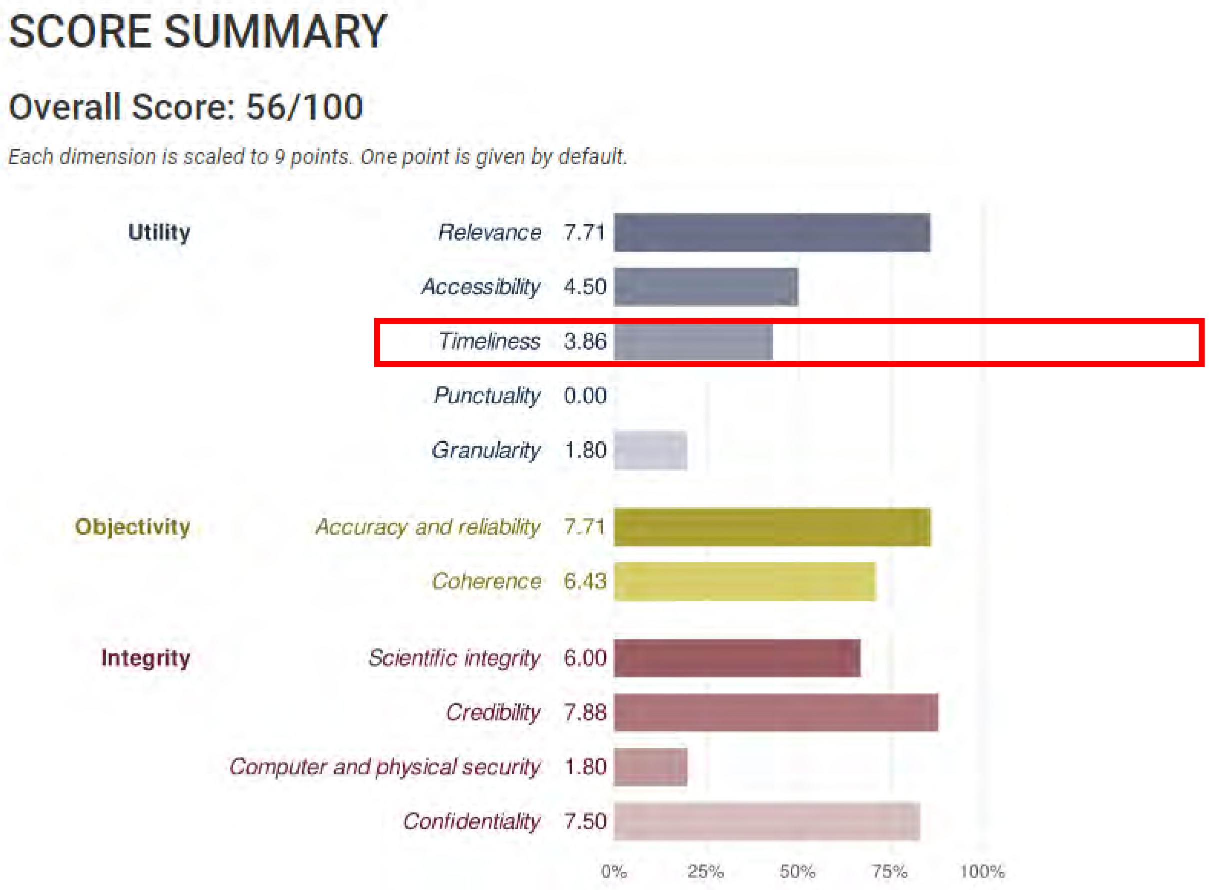
SCORE SUMMARY

Overall Score: 56/100

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

Discussion

Consider use case: Some data sources have advantages with accuracy, **timeliness**



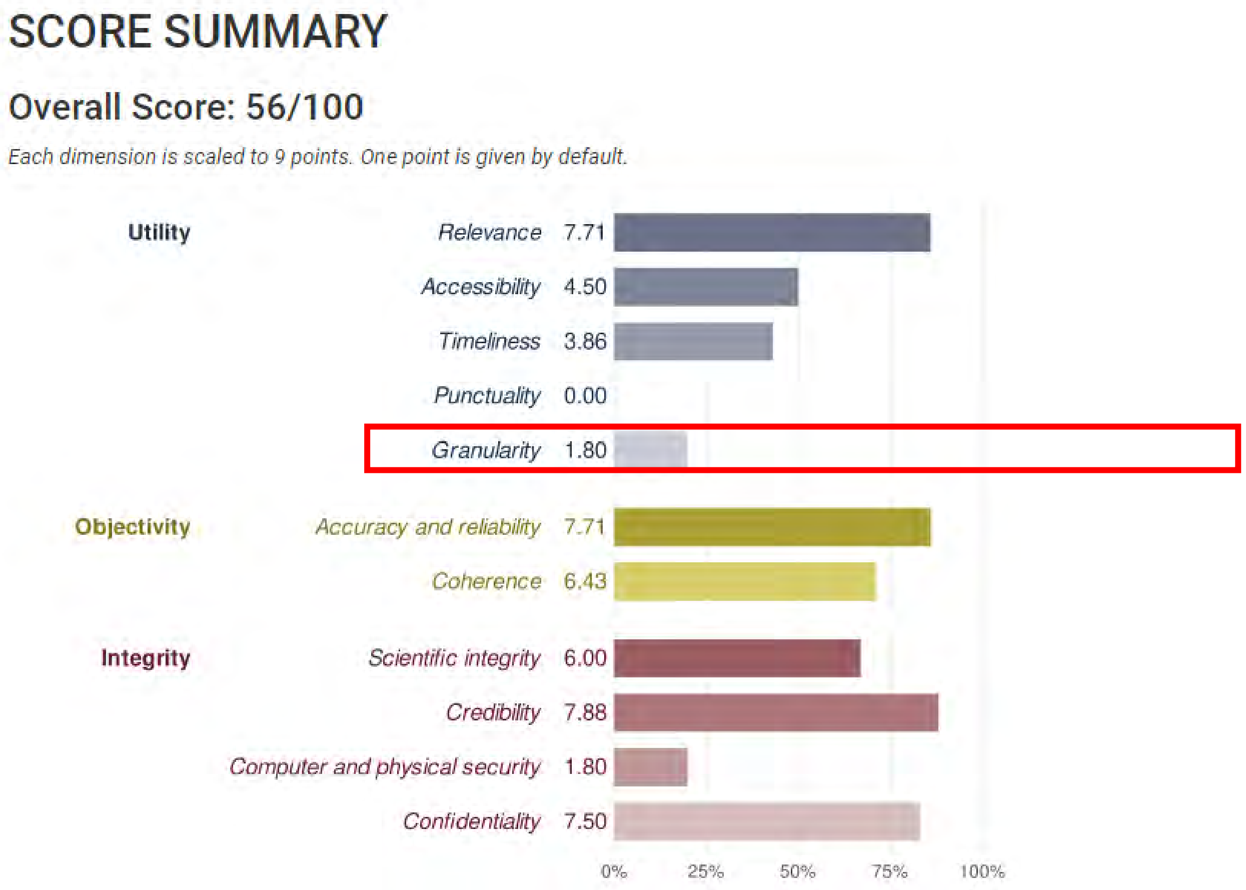
SCORE SUMMARY

Overall Score: 56/100

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

Discussion

Consider use case: Some data sources have advantages with accuracy, timeliness, and **granularity**



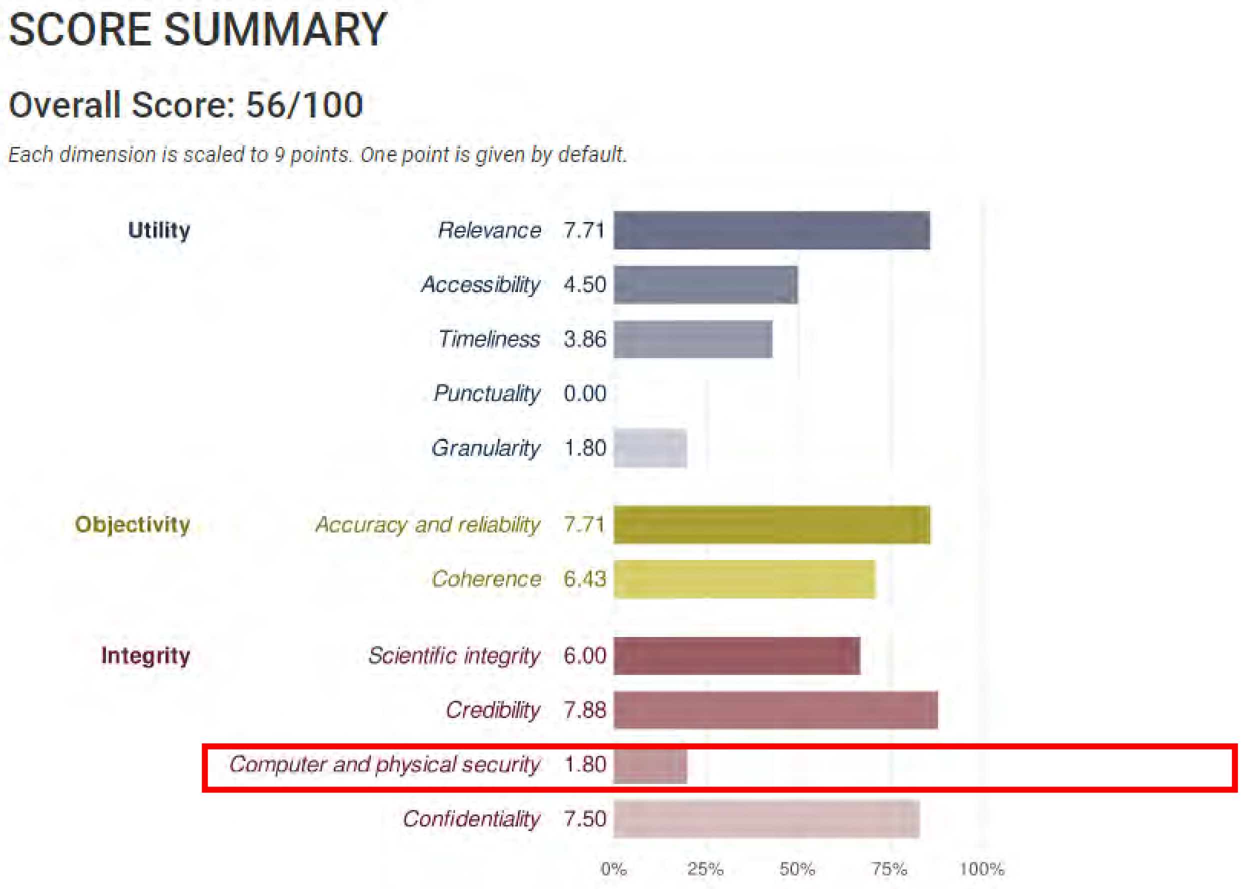
SCORE SUMMARY

Overall Score: 56/100

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

Discussion

Security documentation may be sparse



SCORE SUMMARY

Overall Score: 56/100

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

Discussion

The scorecard allows for ready and systematic data quality assessment of a range of data files applying the FCSM Data Quality Framework.

- Yes/no questions support scoring by data quality dimensions and comparing data files
- Tool incorporates technological advantages of R Shiny and R Markdown implementation and report generation
- Provides a way to quantify qualitative measures

Elizabeth Mannshardt

emannsha@nsf.gov

https://ncses.nsf.gov

@NCSESgov