Federal Committee on Statistical Methodology
Work Group on Transparent Reporting of Data Quality

A Framework for Data Quality

Rolf R. Schmitt, PhD
Bureau of Transportation Statistics

November, 2020

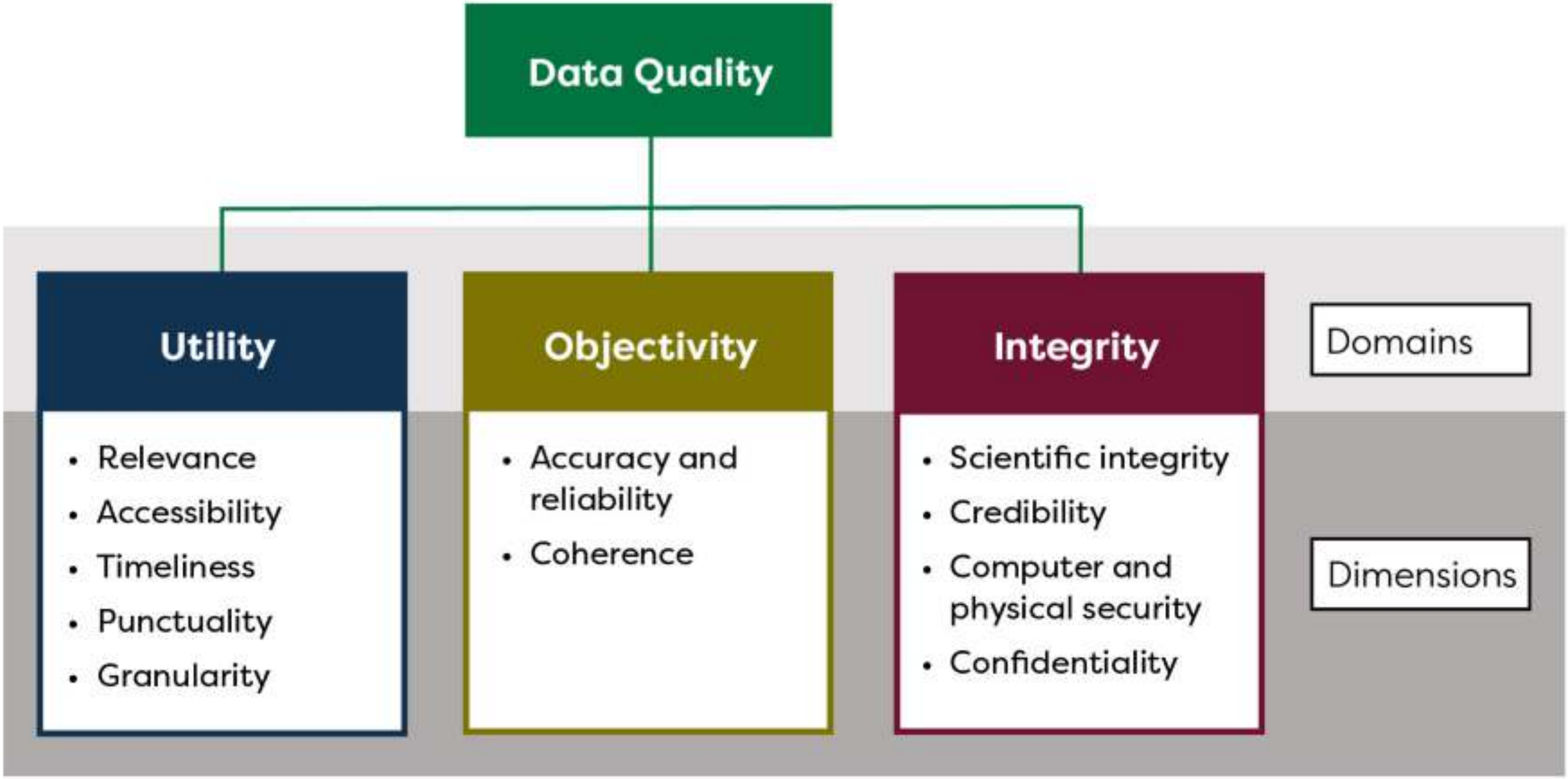
A brief history

- Chamberlin on data quality tradeoffs
- Program evaluation research
- Total survey error
- The Information Quality Act of 2001
- The Evidence Act, Federal Data Strategy, CNStat research on blended data
- ICSP policy and charge to FCSM
- FCSM workshops and preliminary report

A Framework for Data Quality

- Builds on experience of the Federal Statistical System
- Explains for a broad audience the importance of understanding data quality to determine fitness for purpose
- Organizes the many elements of data quality around the structure of the Information Quality Act
- Provides strategies for documenting and reporting data quality

Organizing threats to data quality



Data Quality

Utility

- Relevance
- Accessibility
- Timeliness
- Punctuality
- Granularity

Objectivity

- Accuracy and reliability
- Coherence

Integrity

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

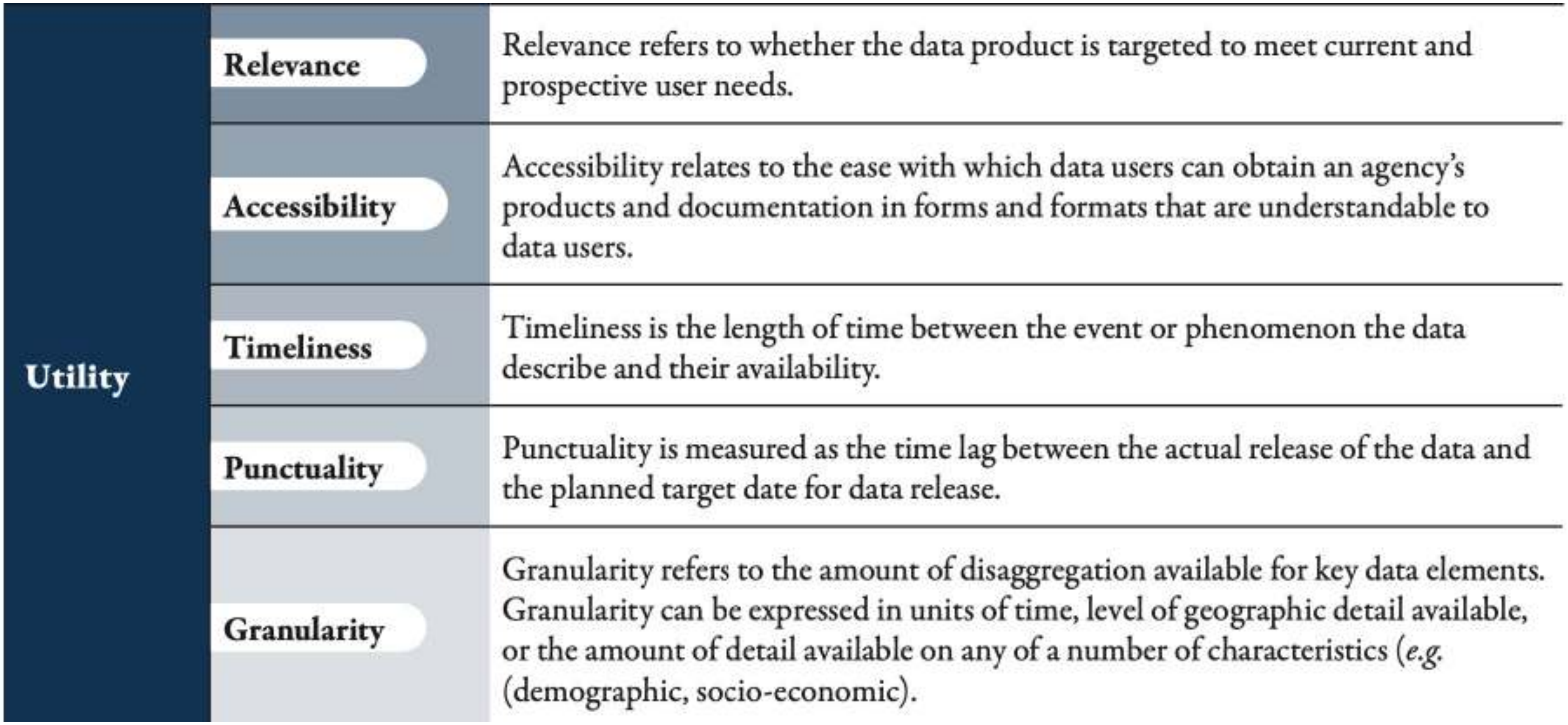
Domains

Dimensions

Factors that affect data quality

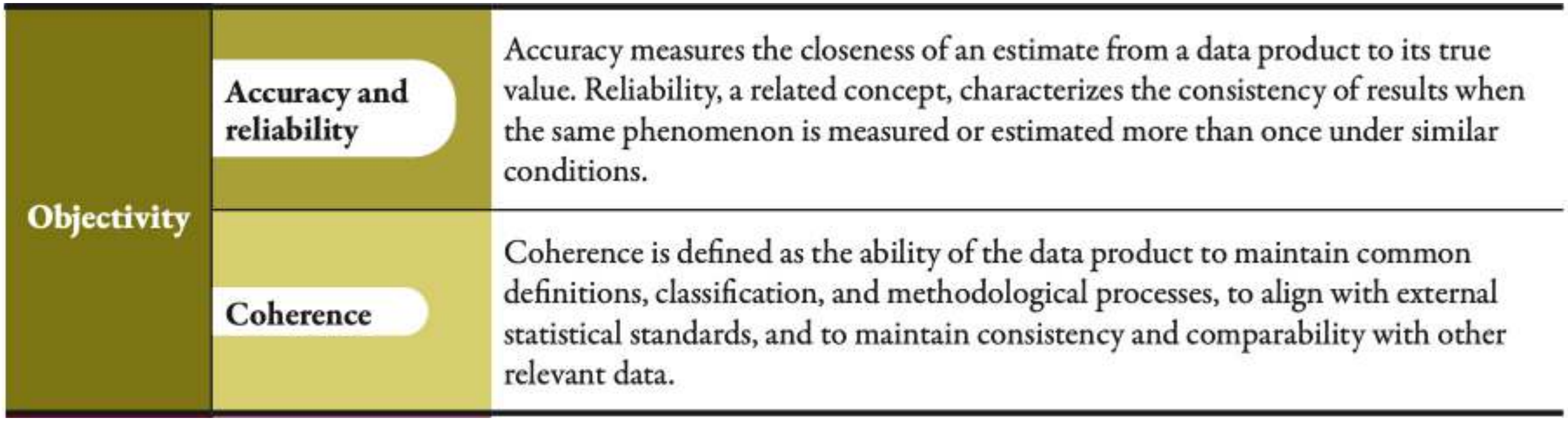
- Provides an inventory of threats to each dimension of data quality and examples of ways to manage the threats
- Covers all types of data
  - Surveys
  - Administrative records
  - Sensor data
  - Integrated/blended data and estimates
- Touches on special topics such as geographic data

Organizing threats to data quality



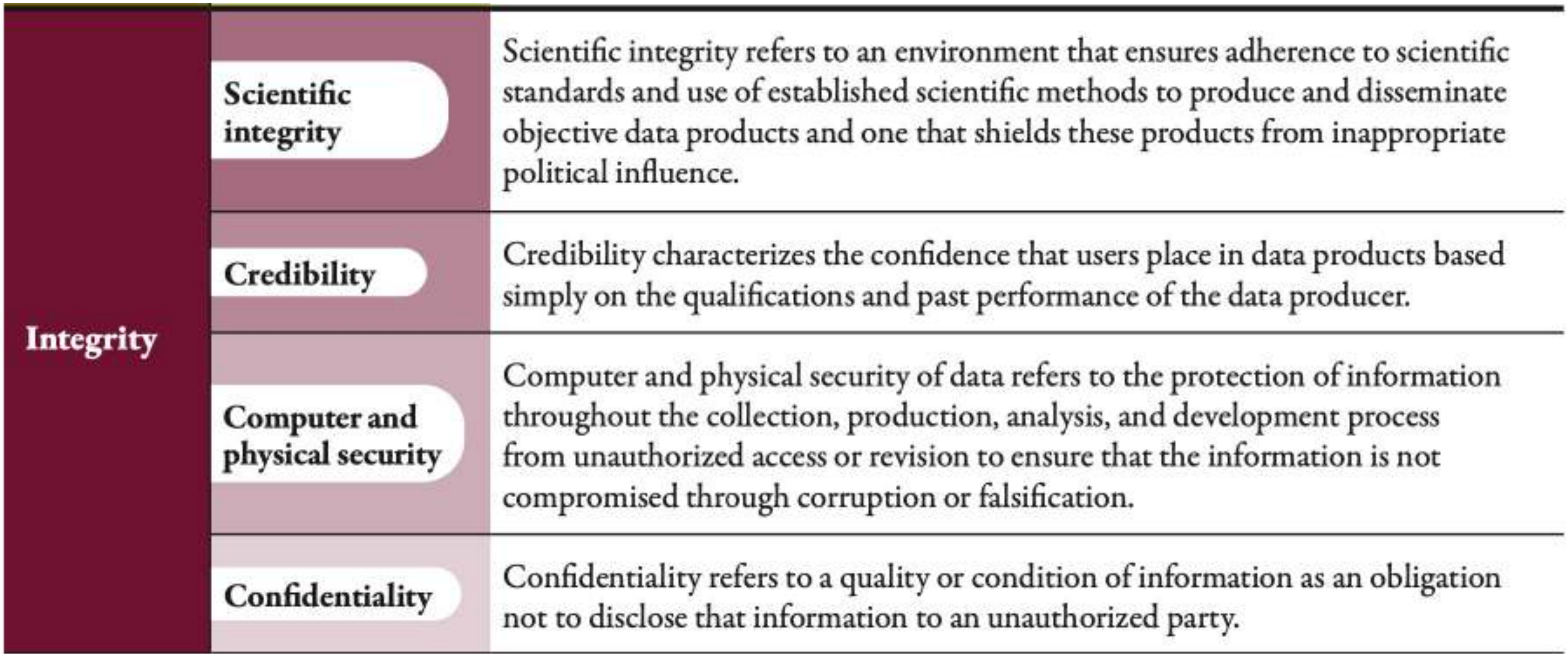
Utility	Relevance	Relevance refers to whether the data product is targeted to meet current and prospective user needs.
	Accessibility	Accessibility relates to the ease with which data users can obtain an agency's products and documentation in forms and formats that are understandable to data users.
	Timeliness	Timeliness is the length of time between the event or phenomenon the data describe and their availability.
	Punctuality	Punctuality is measured as the time lag between the actual release of the data and the planned target date for data release.
	Granularity	Granularity refers to the amount of disaggregation available for key data elements. Granularity can be expressed in units of time, level of geographic detail available, or the amount of detail available on any of a number of characteristics (*e.g.* (demographic, socio-economic).

Organizing threats to data quality



Objectivity	Accuracy and reliability	Accuracy measures the closeness of an estimate from a data product to its true value. Reliability, a related concept, characterizes the consistency of results when the same phenomenon is measured or estimated more than once under similar conditions.
	Coherence	Coherence is defined as the ability of the data product to maintain common definitions, classification, and methodological processes, to align with external statistical standards, and to maintain consistency and comparability with other relevant data.

Organizing threats to data quality



Integrity	**Scientific integrity**	Scientific integrity refers to an environment that ensures adherence to scientific standards and use of established scientific methods to produce and disseminate objective data products and one that shields these products from inappropriate political influence.
	Credibility	Credibility characterizes the confidence that users place in data products based simply on the qualifications and past performance of the data producer.
	Computer and physical security	Computer and physical security of data refers to the protection of information throughout the collection, production, analysis, and development process from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification.
	Confidentiality	Confidentiality refers to a quality or condition of information as an obligation not to disclose that information to an unauthorized party.

With respect to the quality of integrated/blended data

- Quality of integrated/blended data is often different than the sum of the qualities of its parts
  - Offsetting versus exacerbating quality problems
- Estimating the aggregate quality
  - Comparisons of competing estimates
  - Sensitivity analysis

Reporting data quality

- Avoids past emphasis on large and resource-consuming data quality profiles
- Applies to managers of data collection programs and to analysts
- Three audiences
  - The data program manager / analyst
  - The power user
  - The occasional user or decisionmaker

Reporting data quality

- The cultural change for program managers and analysts: consider all threats and note how you address each relevant threat to inform your successor
- The manager's notes provide a cornerstone for technical documentation for power users
- The elevator speech: describe in a few words how likely the data will misguide a decision

Lessons and future work

- The Information Quality Act provides a useful framework for examining data quality
- Data quality tradeoffs change over time
  - Covid-19 put a premium on timeliness over deliberative vetting of accuracy
- More work is needed
  - Strategies for developing and updating data quality standards
  - Additional tools to measure quality in blended data sets
  - Best practices for identifying quality of data obtained from sources that lack transparency and from advanced (AI) algorithms
  - Tools for harvesting data quality notes into metadata and into effective caveats for power users
  - Effective labeling of carefully vetted data versus experimental data
  - Communicating data quality while building trust
  - Other …

Conclusion

- All data have problems, but do the problems matter for the decision at hand?
- Data managers should consider all possible data quality problems, deal with problems that can reasonable be addressed, and document how they dealt each problem for their successors
- Include data quality in guides for power users and summarize the problems for an elevator speech to tell occasional users how far they can take the data without misguiding decisions that have important consequences

For the details

- The full report is available at: https://nces.ed.gov/fcsm/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf
- Information about the Federal Committee on Statistical Methodology is posted at: https://nces.ed.gov/fcsm/