Leveraging the Data Quality Framework Workshop

September 10, 2021

A Virtual Workshop Sponsored by the Interagency Council on Statistical Policy, Chief Data Officer Council, and Evaluation Officer Council
Hubert Hamer, USDA NASS

Opening Remarks
Mentimeter

• Go to https://www.menti.com/xau2spxsrs

• Go to menti.com and enter the code: 4081 7831

• Scan the QR code with your mobile device
The FCSM Framework for Data Quality

Jennifer D. Parker, Ph.D.
National Center for Health Statistics

Leveraging the Data Quality Framework
September 10, 2021
Background

• Understanding data quality is essential for data-driven decision making
  – Data users who understand the “fitness-for-use” of data products are more likely to use them appropriately
  – Higher-impact uses of data require higher quality data
• All data have strengths and weaknesses
• Data quality for surveys is relatively well-established but data quality for integrated data and other non-statistical data are less developed
Data Quality Milestones 2001-2020 (A)

WP #31
- 2001 Measuring and Reporting Sources of Error in Surveys
- Focus on reporting accuracy of survey data outputs

IQ Act
- 2001 Information Quality Act/OMB Guidelines
- Provided a framework, with a call for more detailed OMB and Agency Guidelines

Standards
- Emphasis on survey data quality
Data Quality Milestones 2001-2020 (B)

**Integrated Data**
- System-wide declining response rates, increasing costs
- Increased use of non survey data sources, alone or integrated with statistical survey data

**Evaluations**
- 2015-17 Two CNSTAT reports and the Commission on Evidence-Based Policymaking, integrated data
- New visions for Federal statistics; identified obstacles and provided recommendations for moving forward

**2018 Evidence Act**
- Federal Data Strategy, Foundations for Evidence Based Policymaking Act, revised OMB Information Quality Act Guidelines
- Address data quality and compatibility with integrated data

**Quality Framework**
- 2018-2019 seminars on data quality and integrated data
- FCSM Data Quality Framework
FCSM Framework for Data Quality

• Builds on experience of the Federal Statistical System
• Organizes the elements of data quality around the structure of the Information Quality Act
• Explains for a broad audience the importance of understanding data quality to determine fitness for purpose, identifying and mitigating key data quality threats, and evaluating trade-offs
• Provides strategies for documenting and reporting data quality
FCSM Framework for Data Quality

- **Utility**
  - Relevance
  - Accessibility
  - Timeliness
  - Punctuality
  - Granularity

- **Objectivity**
  - Accuracy and reliability
  - Coherence

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

Domains

Dimensions
Domains of Data Quality

- **Utility** - the extent to which information is well-targeted to valuable needs: it reflects the usefulness of the information to the intended users.
- **Objectivity** - whether information is accurate, reliable, and unbiased, and is presented in an accurate, clear and interpretable, and unbiased manner.
- **Integrity** – the maintenance of rigorous scientific standards and the protection of information from manipulation or influence as well as unauthorized access or revision.
Dimensions of Utility - I

- **Relevance**: whether the data product is targeted to meet current and prospective user needs
- **Credibility**: the confidence that users place in data products based simply on the image of the data producer
- **Accessibility**: the ease with which data users can obtain an agency’s products and documentation in forms and formats that are understandable to data users.
Dimensions of Utility - II

- **Timeliness**: the length of time between the event or phenomenon the data describe and their availability
- **Punctuality**: the time lag between the actual release of the data and the planned target date for data release
- **Granularity**: the amount of disaggregation available for key data elements.
Dimensions of Objectivity

• **Accuracy**: the closeness of an estimate from a data product to its true value
  – **Reliability**: characterization of repeated estimates of accuracy over time

• **Coherence**: 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
Dimensions of Integrity

- **Scientific Integrity**: an environment that ensures the use of established scientific methods to produce and disseminate objective data products and shields these products from inappropriate political influence.

- **Computer and Physical Security**: 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**: a quality or condition of information as an obligation not to disclose that information to an unauthorized party.
Threats to Data Quality

• Threats can be identified for all dimensions
  – Threats can be relevant for multiple dimensions
  – Mitigating threats for one dimension can increase threats for another

• Managing trade-offs among quality dimensions is important

• Threats to quality for blended data combine threats for data inputs, blending methods, and data outputs
Assessing Data Quality

• Regularly identify threats to data quality for ongoing data collections, including when considering new source data for inclusion
  – Decisions on trade-offs among threats and mitigation measures should be considered in the context of the data’s purpose and all identified threats
  – Data quality for the intended use may differ from that for its original purpose
Conclusion

• Data quality has been long studied for statistical data, especially surveys, but is less developed for integrated and secondary-use data

• The FCSM Data Quality Framework can be used to evaluate quality for all data
Using the Framework for Data Quality

Rolf R. Schmitt, PhD
Bureau of Transportation Statistics

September 2021
The Framework for Data Quality

• Organizes the many elements of data quality around the structure of the Information Quality Act
• Provides a comprehensive and consistent terminology to describe the many aspects of data quality
• Looks overwhelming to use and burdensome to report
Don’t panic

• Many data quality threats can be dismissed after brief consideration for a data program
• There are few universal rules for weighing importance of one data quality concern over another: tradeoffs are expected
• Documentation while planning and doing what you do is a good habit that helps your successors and supports transparency
Reporting data quality

• Data quality reports as a byproduct of documenting your work
• 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
Tradeoffs change over time

• Covid-19 put a premium on timeliness over deliberative vetting of accuracy

• “It may be better, in the gross affairs of life, to be less precise and more prompt. Quick decisions, though they may contain a grain of error, are often better than precise decisions at the expense of time.”
  – T.C. Chamberlin, President of the University of Wisconsin, 1890
Future work

• 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
Conclusion

• By using the structure and terminology of the Framework, we will have a common basis for sharing information about data quality across agencies and with the public

• A common language will support transparency about our current data and analyses and a common basis for considering improvements in data and analysis
For the details

DESIGNING A DATA QUALITY POLICY
NSF’s Frame of Reference

Evidence Act

Learning Agendas: “Systematic way to identify the data agencies intend to collect, use, or acquire, as well as the methods and analytical approaches to facilitate the use of evidence in policymaking”

Federal Data Strategy

The CDO shall “ensure that, to the extent practicable, the agency maximizes the use of data in the agency”

Mission

To promote the progress of science; to advance the national health, prosperity, and welfare; and to secure the national defense; and for other purposes.

(Mission

(NSF Foundation Act, 1950)
Leveraging the Data Quality Framework

“It applies to all data: data collections and data systems; restricted and public use microdata files; data products produced through data integration, modeling, harmonization and other statistical analyses; and analysis outputs, such as tables, estimates, graphics and reports.” FCSM Data Quality Framework

To be a strategic asset, data must be transparent, verified, and documented…

across the data, information, knowledge stages.
Dimensions of the Data Quality Framework

- **Utility**
  - Relevance
  - Accessibility
  - Timeliness
  - Punctuality
  - Granularity

- **Objectivity**
  - Accuracy and reliability
  - Coherence

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

Domains:

- Dimensions
ATLAS Experiment at CERN
To be a **strategic asset**, data must be **transparent, verified, and documented** across the **data, information, knowledge** stages.
Data Quality Policy

Data Inventory

Data management lifecycle is standardized, accessible and detailed

Data Inventory Standards

Analytic Tools

Tools processing the data are documented and vetted

Analytic Tool Standards

Analytic Outputs

Queries and analyses are documented and reproducible

Best Practices for Analytics Documentation
Data Quality Policy - Process

**Data Inventory**
- Inventories prepared by Data Stewards
  - Feedback provided by internal user community
  - Reviewed and approved by Data Governing Body

**Analytic Tools**
- Tools used at Enterprise level
  - Methods and documentation reviewed by Data Governing Body

**Analytic Outputs**
- Output generated by office
  - Documented and archived internally by office standards
Thank You!

Avital Percher: apercher@nsf.gov
Dorothy Aronson (CIO/CDO): daronson@nsf.gov
# Data Inventory Standards – Objectives

Defines metadata documentation standards and review process

<table>
<thead>
<tr>
<th>Compliance</th>
<th>Roles &amp; Responsibilities</th>
<th>Documentation Maintenance</th>
<th>Master Metadata Schema</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Agency compliance with federal mandates</td>
<td>Define the roles and responsibilities in the Data Inventory Management Process</td>
<td>Define the requirements for maintaining metadata profiles and data dictionaries of NSF’s data repositories</td>
<td>Define a master metadata and dictionary schema as an agency standard</td>
<td>Define a user inclusive validation process</td>
</tr>
</tbody>
</table>
## Analytic Tools Standards - Objectives

Defines documentation requirements and validation process for tools used on an ‘enterprise’ level.

<table>
<thead>
<tr>
<th>Community Standards</th>
<th>Tool Documentation Benchmark</th>
<th>Review and Approval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Define a community standard of excellence and support leadership’s need for trustworthy and vetted data tools.</td>
<td>Establish a benchmark for tool documentation to promote development and application practices that align with community best practices.</td>
<td>Describe a review and approval process by the EADGSC to support the NSF community’s need for tools vetted by data experts.</td>
</tr>
</tbody>
</table>
## Best practices for analytics documentation - Objectives

Defines guidelines for documenting analytics outputs

<table>
<thead>
<tr>
<th>Improved Quality Standards</th>
<th>Replication</th>
<th>Knowledge Dissemination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhance the quality and trustworthiness of the data collection and analysis.</td>
<td>Enable replication of the analysis as needed in the future, by both the office and others.</td>
<td>Allow the adaptation of the study to other needs of the community, increasing efficiency.</td>
</tr>
</tbody>
</table>
Data Inventory Standards

EDI RPTSQL PILOT STAGES

STAGE 1

Collect data lineage information from technical data stewards

Output
Draft metadata & data dictionary for RPTSQL tables

STAGE 2

Validate data with domain data steward expertise and submit for Data Governance (EAGDSC) approval

Output
Finalized metadata & data dictionary Inputs for master data management

STAGE 3

Publish validated and approved data for internal NSF use

Output
Published & searchable data inventory (metadata & data dictionary)

1Stage 1 corresponds with the beginning of Q3. Activities preceding Q3 are not included in the listed stages.
Quality Considerations for Alternative Data: A Case Study using CORP5 Data

John Bieler
Senior Economist, CPI
September 10, 2021
Familiar?

Quality Considerations for Alternative Data in the BLS Producer and Consumer Price Indexes

Crystal Konny (CPI)
Bonnie Murphy (PPI)

December 2017
Scorecard for Alternative Data

<table>
<thead>
<tr>
<th>Quality Metrics</th>
<th>Sample Frames</th>
<th>Benchmarking</th>
<th>Hedonics</th>
<th>Replace collection</th>
<th>Supplement Collection</th>
<th>Data Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Transparency- methods understood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Granularity- Level of detail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality of descriptive data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scope, type of price</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage- items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage- geography</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage- outlets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling procedures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data delivery reliable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Viability of data source</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Usability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Frequency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Security</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data delivery timeliness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data history</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Cleanliness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Usability- mods to current system</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The 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
Background on CORP5

- CORP5 is a secondary source of gas price data
- Average of roughly 205,000 reported gas price observations every day
  - Roughly 6.23 million gas prices every month!
- Gas prices are updated in real-time
- CPI receives data the following day
- CORP5 data includes prices for three categories: Regular unleaded gasoline, Mid-grade, and Premium
- BLS obtained approval from CORP5 to use their data and began to voluntarily provide their data using a secure portal
<table>
<thead>
<tr>
<th>Domain</th>
<th>Dimension</th>
<th>Definition</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility</td>
<td>Relevance</td>
<td>Relevance refers to whether the data product is targeted to meet current and prospective user needs.</td>
<td>Is the data a relevant input to our data products and measurement our measurement objective?</td>
<td>CORP5 provides daily gasoline prices for thousands of stations across the U.S. Produce indexes and average prices for gasoline and individual fuel types.</td>
</tr>
<tr>
<td>Utility</td>
<td>Accessibility</td>
<td>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.</td>
<td>Are the costs to access the data an effective use of resources? Will the methodology limit our ability to release data to users? How can we describe the methodology to data users?</td>
<td>CORP5 is providing the data on a voluntary basis. Make a public announcement in advance and provide materials online, such as factsheets and articles.</td>
</tr>
</tbody>
</table>
Transmission of material in this release is embargoed until
8:30 a.m. (ET) April 13, 2021

Technical information: (202) 691-7000 • cpi_info@bls.gov • www.bls.gov/cpi
Media Contact: (202) 691-5902 • PressOffice@bls.gov

CONSUMER PRICE INDEX – MARCH 2021

The Consumer Price Index for All Urban Consumers (CPI-U) increased 0.6 percent in March on a
seasonally adjusted basis after rising 0.4 percent in February, the U.S. Bureau of Labor Statistics
reported today. The March 1-month increase was the largest rise since a 0.6-percent increase in August
2012. Over the last 12 months, the all items index increased 2.6 percent before seasonal adjustment.

The gasoline index continued to increase, rising 9.1 percent in March and accounting for nearly half of
the seasonally adjusted increase in the all items index. The natural gas index also rose, contributing to a
5.0-percent increase in the energy index over the month. The food index rose 0.1 percent in March, with
the food at home index and the food away from home index both also rising 0.1 percent.
## CORP5 case study cont.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dimension</th>
<th>Definition</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility</td>
<td>Timeliness</td>
<td>Timeliness is the length of time between the event or phenomenon the data describe and their availability.</td>
<td>Are the data representative of the index reference period?</td>
<td>Yes, daily prices across the month.</td>
</tr>
<tr>
<td>Utility</td>
<td>Punctuality</td>
<td>Punctuality is measured as the time lag between the actual release of the data and the planned target date for data release.</td>
<td>Can the methodology be implemented within the typical production processing schedule?</td>
<td>Yes, CORP5 will be implemented into the current production schedule. We are currently parallel testing.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>What is the probability and impact on the production schedule due to delayed delivery of data or unexpected time needed to process data?</td>
<td>Based on multi year evaluation period, the probability of an impact is low.</td>
</tr>
</tbody>
</table>
## Corp5 case study cont.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dimension</th>
<th>Definition</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility</td>
<td>Granularity</td>
<td>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).</td>
<td>Is there adequate data to support the current level of granularity in data products?</td>
<td>Yes, we will produce price indexes and average price products at the same level of granularity.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Is there sufficient data to adequately protect confidentiality?</td>
<td>Yes, thousands of gas stations protecting confidentiality.</td>
</tr>
</tbody>
</table>
Comparison of regular gas prices

CPI
Gas prices (reg. grade) - November 6, 2019 in CPI samples

CORP 5
Gas prices (reg. grade) - November 6, 2019
Comparing number of prices

<table>
<thead>
<tr>
<th></th>
<th>CORP5 Daily Price Observations</th>
<th>CORP5 Monthly Average Prices</th>
<th>CPI Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2,516,768</td>
<td>139,107</td>
<td>4,500</td>
</tr>
</tbody>
</table>
## CORP5 case study cont.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dimension</th>
<th>Definition</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectivity</td>
<td>Accuracy and reliability</td>
<td>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.</td>
<td>Any concerns with the qualitative assessment of total measurement error?</td>
<td>No, research results compared favorably to the CPI Gasoline index at the U.S. Level.</td>
</tr>
<tr>
<td>Coherence</td>
<td>Coherence</td>
<td>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.</td>
<td>Does the methodology impact the ability to compare CPI data with external sources?</td>
<td>No, the methodology is still comparable with external sources.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Is the methodology coherent with other CPI methodologies (not just what it is replacing)?</td>
<td>Yes, a mix of geomeans and Laspeyres index methodology. Added additional aggregation steps.</td>
</tr>
</tbody>
</table>
CORP5 Research – Differences never greater than 1% at U.S. level for gasoline
### CORP5 Stored Meta Data

<table>
<thead>
<tr>
<th>AREA</th>
<th>AREA_DESC</th>
<th>NUM_PR_OBS</th>
<th>NUM_RELATIVES</th>
<th>NUM_PHYS_LOCATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>U.S.</td>
<td>9,576,611</td>
<td>129,755</td>
<td>50,049</td>
</tr>
<tr>
<td>N000</td>
<td>Non-Self-Representing PSUs</td>
<td>3,745,235</td>
<td>42,976</td>
<td>16,753</td>
</tr>
<tr>
<td>S000</td>
<td>Self-Representing PSUs</td>
<td>5,831,376</td>
<td>86,779</td>
<td>33,296</td>
</tr>
<tr>
<td>S12A</td>
<td>New York-Newark-Jersey City, NY-NJ-PA</td>
<td>265,885</td>
<td>10,187</td>
<td>3,932</td>
</tr>
<tr>
<td>S23A</td>
<td>Chicago-Naperville-Elgin, IL-IN-WI</td>
<td>901,102</td>
<td>6,855</td>
<td>2,462</td>
</tr>
<tr>
<td>S49A</td>
<td>Los Angeles-Long Beach-Anaheim, CA</td>
<td>447,457</td>
<td>7,086</td>
<td>2,408</td>
</tr>
</tbody>
</table>
## CORP5 case study cont.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dimension</th>
<th>Definition</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrity</td>
<td>Scientific integrity</td>
<td>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.</td>
<td>What is the probability and impact of the data provider (either maliciously or unintentionally) interfering with the data in a way that impacts estimates?</td>
<td>The probability is low and the impact is low. There is no incentive for the data provider to manipulate the data.</td>
</tr>
<tr>
<td>Credibility</td>
<td></td>
<td>Credibility characterizes the confidence that users place in data products based simply on the qualifications and past performance of the data producer.</td>
<td>Review the output of index simulations. The more a simulation deviates from production, the more of an understanding approvers would like to have of the cause of differences.</td>
<td>Often cited source in news organizations and widely accepted by users as a credible source of price information.</td>
</tr>
</tbody>
</table>
## CORP5 case study cont.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dimension</th>
<th>Definition</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrity</td>
<td>Computer and physical security</td>
<td>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.</td>
<td>What is the probability and impact of risks of a loss of data or data quality issues due to technical issues?</td>
<td>The fallback plan is to use CPI collected data.</td>
</tr>
<tr>
<td>Domain</td>
<td>Dimension</td>
<td>Definition</td>
<td>Question</td>
<td>Answer</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------</td>
<td>----------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Cost effectiveness</td>
<td></td>
<td></td>
<td>Are the new data and methods cost effective relative to the data and methods they are replacing? Include development costs, contracting costs, data collection costs, data storage, and maintenance costs.</td>
<td>Using the CORP5 data is cost neutral at this point.</td>
</tr>
</tbody>
</table>
Contact Information

John Bieler
Senior Economist
Consumer Price Index
(202) 691-5407
bieler.john@bls.gov
Data Quality Evolution

Dorothy Aronson
Chief Information Officer/Chief Data Officer
09/10/21
When I think about Data Quality for NSF...

Moving to the cloud has allowed agencies to store and use massive amounts of data more efficiently than when they maintained their own data centers. Now the next big step for many of those agencies, especially if their goal is to implement emerging technologies like artificial intelligence, is figuring out how to get the most out of that data. For a number of federal chief data officers and data experts, that means improving data quality.

Data quality is important to these efforts, and that doesn’t just mean ensuring the numbers in a table are accurate. Utility and trustworthiness of data is also determined by documenting where the data comes from, how often it’s updated, and who has access to it. That’s why some CDOs are working on how to establish a stronger foundation or framework to ensure processes and data management are meeting specific standards in order to facilitate easy data sharing.

Source: Data quality, framework, accessibility are key to implementing emerging technologies | Federal News Network
When I think about Data Quality for NSF...
When I think about Data Quality for NSF...
When I think about Data Quality for NSF...
When I think about Data Quality for NSF...

What does “data quality” even mean?

Scientific Integrity
Confidentiality
Relevance
Coherence
Accuracy
Accessibility
Credibility
Reliability
Computer and Physical Security
Punctuality
Granularity
Timeliness

CIO
CDO
FCSM offers a framework.

What does “data quality” even mean?
NSF aligns within the framework.
BLS demonstrates alignment with the framework.
Framework provides a common language...

...allowing necessary variation to fit mission.
NSF’s Data Quality: Lessons Learned

- Center the end user
- Bottom-up and top-down tactics
- Underscore agency goals and align with federal policies
- Focus on progress over perfection
## NSF’s Data Quality Initiatives: Challenges and Solutions

<table>
<thead>
<tr>
<th>ISSUE</th>
<th>CHALLENGE</th>
<th>SOLUTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Getting Started</td>
<td>When creating a policy from scratch there is significant time spent collecting artifacts.</td>
<td>Use the numerous existing resources to create the basis for the policy.</td>
</tr>
<tr>
<td>Maintaining Scope</td>
<td>Through the drafting and review process several other policy needs were identified.</td>
<td>Instead of incorporating the ideas into the draft policy, log the ideas for future policy development efforts.</td>
</tr>
<tr>
<td>Establishing an Inclusive Process</td>
<td>Numerous stakeholders have an interest in supporting the development of the policy.</td>
<td>Small teams assist in policy development. Iterative and inclusive review process. Tailored briefings for senior staff.</td>
</tr>
<tr>
<td>Making the Change Stick</td>
<td>Implementing a new policy requires buy-in across the agency.</td>
<td>(In process) Imbed Data Governance Group in policy implementation. Work to build a policy and tools that provide value to stakeholders.</td>
</tr>
</tbody>
</table>
BLS’ CORP5 Case Study: Standout Points

- Integrating existing frameworks and resources (e.g., The Framework and the Scorecard for Alternative Data)

- Showcasing scaled impact across the alternative data sources

- Strengthening evidence-building efforts by using secondary data sources to supplement agency data
Contact Information

Dorothy Aronson
CIO/CDO
NSF
daronson@nsf.gov
703.292.4299