Principles for Modernizing Production of Federal Statistics,
Interagency Council on Statistical Policy

Background

The Federal Statistical System has a long history of collecting data to produce high quality official statistics. Most official statistics are based on surveys or highly curated administrative data, allowing the agencies to control measurement error and transparently report data limitations to the public as required by U.S. Office of Management and Budget (OMB) standards. Increasingly, the Federal Statistical System is facing challenges maintaining statistical surveys, primarily due to falling response rates and rising costs. At the same time, advances in technology and methodology are creating opportunities for statistical agencies to modernize their practices. As noted in a recent set of reports from the National Academies of Sciences, Engineering and Medicine (see Appendix) a major component of modernization is shifting reliance from primarily high quality yet expensive and burdensome surveys to the acquisition and curation of data not initially designed for statistical purposes. This non-statistical information, whether from within the Federal Government or from external sources, can be used to create statistical information, often by integrating data from multiple sources.

A shift from reliance on surveys for primary data collection to reliance on surveys as a complement to already existing data, either in-house or otherwise available to an agency, is essential to the modernization of the Federal Statistical System. This shift is underway in many agencies and statistical programs. In addition, due to open data initiatives and technological advances, more agencies are providing information or data that statistical agencies and units could use in the creation of statistical information or data.

The use for statistical purposes of non-statistical or integrated data—also known a blended, hybrid, or combined data—requires investment in data acquisition and development of standards. First, statistical agencies and components must be able to access alternative data sources. OMB has previously issued guidance clarifying all agencies' responsibility to steward and make accessible non-statistical, “administrative” data resources for statistical purposes, to the extent permitted by law (see Appendix). Useful data, including sensor data, imagery, web scraping, and outputs of models, to name a few, may be obtained from other sources, including commercial, creating potential concerns for longitudinal affordability and sustainability. A second major challenge is measuring and transparently communicating the quality of statistical information derived from non-statistical or integrated statistical and non-statistical information.

However, non-statistical data are not designed for inference and often fall short of standards set by OMB in Statistical Policy Directive 2 (see Appendix). The population represented in non-statistical data may be uncertain or may be a subset of that desired for official statistics. For example, depending on the source, data items may be defined to facilitate program administration, commercial transactions, marketing needs, etc., and may, therefore, be less than ideal for statistical inferences. Often meta- and para-data are inadequate.

Quality and Transparency

Adhering to long-standing principles on utility, burden, quality, and other data characteristics, while transforming the methods used to produce statistical data products, is important to maintaining public confidence in the integrity of Federal statistics. All data have potential errors, and errors can be compounded when data from different sources are integrated to produce statistical estimates. Poorly estimated or overextended statistics can misguide decision makers, leading to costly consequences. Federal statistical agencies must ensure that non-statistical or integrated data sources result in quality statistics and clearly, meaningfully, and effectively communicate the limitations of those statistics so they are used wisely.

The Interagency Council on Statistical Policy (ICSP)[1], working with the Federal Committee on Statistical Methodology (FCSM), is developing a data quality framework for statistical use of non-statistical and integrated data. The new framework will capitalize on the ICSP and the FCSM's deep understanding of measurement issues and track record in developing and applying methodological innovations. The FCSM has completed a thorough review of existing frameworks and current practices for providing transparency about the quality of products produced from non-statistical or integrated data. In addition, Federal statistical agencies and the FCSM conduct ongoing research to assess and measure the quality of non-statistical and integrated data.

The following principles for the ICSP work on integrated data are intended to guide the FCSM and to establish priorities for research and ongoing work by statistical agencies to advance the use of non-traditional information and data to produce statistical information. These principles will evolve and be refined to reflect recommendations arising from this research. The work by the ICSP and the FCSM fits more broadly under OMB's development of a Federal Data Strategy as part of the President's Management Agenda. The Data Strategy will include principles for agencies to use in implementing the strategy.

Building on the existing guidance and framework for quality (see Appendix), Federal agencies that are combining statistical and non-statistical data for statistical purposes should strive to adopt these principles. They provide guidance for the data supplier as well as the statistical program. Data users represent a third audience for the principles.

[1] The Interagency Council on Statistical Policy is chaired by the U.S. Chief Statistician at the U.S. Office of Management and Budget (OMB) and is comprised of the heads of 13 principal statistical agencies and one non-principal agency. It is responsible for exchanging information about agency programs and activities; working collaboratively to coordinate Federal statistical practices, and providing advice and counsel to OMB on statistical matters.

Principles for Using Non-Statistical Data for Statistical Purposes

1. **Agencies should employ the highest quality reasonably obtainable sources of information, including non-statistical data sets and derivative information in developing statistical datasets in support of mission activities.**

2. **While fully complying with confidentiality and privacy requirements, agencies should continue to make statistical information created in support of mission activities as granular and timely as practicable and widely accessible.**

3. **Agencies should report transparently on the quality of information they disseminate.**

   For example, similar to reporting on statistical information produced from traditional sources, if there are known deficiencies or limitations in products produced from a non-statistical or integrated data source, these should be clearly articulated.

   Quality is composed of multiple dimensions, each of which should be addressed in transparent public reporting:

   A. ***Objectivity,*** as defined under the Information Quality Act (IQA) guidelines (see Appendix), is an overarching component of quality, in that it focuses on whether the disseminated information is being presented in an accurate, clear, complete, and policy neutral manner, and as a matter of substance, is accurate, reliable, and is statistically unbiased. For integrated data sources, objectivity encompasses the individual data sources, methods to integrate data, and the creation and dissemination of estimates based on integrated data.

   B. ***Accuracy*** is often emphasized in traditional information quality reporting frameworks, such as the total survey error model, and includes examining coverage, measurement, processing, and other errors to assess comparability of statistical estimates, which in the past have often been based on statistical samples, to actual population totals. Accuracy continues to be an important dimension of quality. When statistical information is derived from non-statistical data sources, sufficient information about data definitions, reporting processes, and quality controls should be developed to measure accuracy for statistical purposes.

   C. ***Precision***, a measure of how close two or more measurements of the same statistic are to each other, will be important in helping users interpret results produced from data that may not have been produced using classical statistical sampling methods. Precision affects accuracy especially when differences among measures of the same statistics exceed the intended resolution of the measure.

   D. Protecting the ***confidentiality*** of information and the privacy of subjects from whom data are collected is often required by law, and protection methods can affect the accuracy of the underlying data sets and derivative information. Confidentiality policy and methodology must be reported publicly and, when possible, tools should be provided to allow users to assess the impact of disclosure limitation methods on inferences generated using the data.

E. ***Accessibility*** ensures that information has utility. Agencies must provide sufficient contextual information about contents, attributes, and methods of access, often called metadata. Such metadata should also address changes from preceding disseminations. Agencies should strive to make data available in formats that are discoverable, in open (non-proprietary) formats, and accessible to users with vision or other impairments.

F. ***Relevance,*** consistent with agency mission, to support decision-making by the agency and its stakeholders, is an important dimension of quality, especially in so far as it can require trade-offs with accuracy, timeliness, and granularity, and is often at the heart of determining fitness for purpose.

G. ***Comparability and Coherence*** – where agency mission includes release of statistical output, a given release should be internally coherent and consistent (which includes validating logical and mathematical relationships among data items), should rely on common standards and classifications, and should be comparable over a reasonable period.

H. ***Integrity***, as used in the IQA guidelines, means “the protection of information from unauthorized access or revision, to ensure that the information is not compromised through corruption or falsification.”

4. **Characterization of quality should be both quantitative and qualitative, consistent with available information.** The characterizations should be broad enough to inform the anticipated range of uses of the information.

5. **Judgments used in developing data sets, such as assumptions, defaults, and uncertainties, should be stated explicitly**. This is particularly important when data are comprised of multiple sources. The sensitivity of assumptions should be demonstrated whenever possible with a priority given to user needs.

6. **Agencies should work to adopt common language and framework for reporting on the quality of data sets and derivative information they disseminate.** The focus should be on providing information that will meet user needs, which may vary by product and agency. Drawing on industry standards, where they exist, will improve interoperability of Federal and non-federal data.

Appendix - References

I. Information Quality Act (IQA)

(Included in the Treasury and General Government Appropriations Act for Fiscal Year 2001, Public law 106-554), https://www.fws.gov/informationquality/section515.html

The Information Quality Act requires agencies to ensure and maximize the quality, objectivity, utility, and integrity of statistical and non-statistical information disseminated by Federal agencies. Statistical information is based on data collected primarily for creating official statistics and include statistical surveys and censuses; non-statistical information is derived from data that have been primarily collected for some other purpose (administrative data, private sector data etc.). OMB has issued both general principles and detailed statistical standards to inform agency practice in collecting and disseminating statistical information, based on a body of literature about the dimensions of data quality, decades in the making. OMB also has issued principles and some general policy guidance to inform agency practice in the dissemination of non-statistical information.

Guidelines issued by OMB pursuant to the IQA:
67 FR 8452, February 22, 2002, *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies*, https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/assets/OMB/fedreg/reproducible2.pdf.
Office of Management and Budget Memorandum M-07-24, September 19, 2007, *Updated Principles for Risk Analysis*, https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2007/m07-24.pdf.

These guidelines prohibit agencies from disseminating substantive information that does not meet a basic level of quality. "We recognize that some government information may need to meet higher or more specific information quality standards than others. The more important the information, the higher the quality standards to which it should be held, for example, in those situations involving ''influential scientific, financial, or statistical information'' (a phrase defined in these guidelines).

This concept is often referred to as "fitness for use" or "fitness for purpose." Given the wide variety of public and private purposes to which federal information is put, agencies would find it challenging to anticipate every use. Therefore, it is necessary to describe to the public the quality of a disseminated information resource or dataset so anticipated and other eventual users can make informed choices about whether the information meets their needs.

The guidelines recognize, however, that information quality comes at a cost. Accordingly, the agencies should weigh the costs (for example, costs attributable to agency processing effort, respondent burden, maintenance of needed privacy, and assurances of suitable confidentiality) and the benefits of higher information quality in the development of information, and the level of quality to which the information disseminated will be held.

In the guidelines, OMB defines ‘‘quality’’ as the encompassing term, of which ‘‘utility,’’ ‘‘objectivity,’’ and ‘‘integrity’’ are the constituents. In other guidance (under the Paperwork Reduction Act (PRA) (https://www.opm.gov/about-us/open-government/digital-government-strategy/fitara/paperwork-reduction-act-guide.pdf), OMB elaborates on the definition of quality to include:

A. *Objectivity*
B. *Accuracy*
C. *Accessibility*
D. *Relevance* and *timeliness*
E. *Comparability and Coherence*
F. Transparency about all the above.

II. OMB Statistical Policy Directives (issued as guidance under the PRA)

Statistical Policy Directive #1, 79 FR 71610, December 2, 2014, *Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units* https://www.gpo.gov/fdsys/pkg/FR-2014-12-02/pdf/2014-28326.pdf)

Directive 1 states, that pursuant to the Information Quality Act, OMB has established guidelines that require each Federal agency to institute procedures to ensure the objectivity, utility, and integrity of information, including statistical information, provided to the public. OMB's government-wide Information Quality Guidelines define objectivity, utility and integrity in a manner consistent with use of these terms in the PRA. Each Federal agency, through the adoption or adaptation of these guidelines, maintains its commitment to use the best available science and statistical methods; subjects information, models, and analytic results to independent peer review by qualified experts, when appropriate; disseminates its data and analytic products with a high degree of transparency about the data and methods to facilitate their reproducibility by qualified third parties; and ensures that the presentation of information is comprehensive, informative, and understandable.

Statistical Policy Directive #2:
(71 FR 55522, Sept. 22, 2006 *Standards and Guidelines for Statistical Surveys*) (https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.)

This directive describes specific practices that support the quality of design, collection, processing, production, analysis, review, and dissemination of information from statistical surveys.

III. OMB Guidance on Access to Federal Data Resources for Statistical Purposes

Office of Management and Budget Memorandum M-13-13, May 9, 2013, *Open Data Policy-Managing Information as an Asset,* https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf*.*

This Memorandum establishes a framework to help institutionalize the principles of effective information management at each stage of the information's life cycle to promote interoperability and openness. It requires agencies to collect or create information in a way that supports downstream information processing and dissemination activities, and includes using machine readable and open formats, data standards, and common core and extensible metadata.

Office of Management and Budget Memorandum M-14-06*,* February 14, 2014, *Guidance for Providing and Using Administrative Data for Statistical Purposes*, https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf

This Memorandum provides agencies with guidance for addressing the legal, policy, and operational issues that exist with respect to using administrative data for statistical purposes. It builds on the OMB frame work established in the *Open Data Policy* to help institutionalize the principles of effective information management at each stage of the information's life cycle to promote interoperability and openness.

IV. National Academies of Sciences, Engineering, and Medicine, Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of the Art Estimation Methods

Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy, (2017), Washington, DC, The National Academies Press.
Federal Statistics, Multiple Data Sources, and Privacy Protections: Next Steps, (2017), Washington, DC, The National Academies Press.

These companion reports review current practices in producing Federal statistics, focusing on those that have traditionally relied on surveys as primary sources of data. To mitigate cost and coverage concerns that arise due to the overall decline in survey response rates, the panel explores alternative data sources including federal, state and local administrative data, as well as private sector data. Recommendations focus on developing new infrastructure, methodologies and quality processes to produce statistics from new or integrated data sources. The recommendations also suggest legal and technical frameworks for addressing privacy concerns that arise in using alternate data sources to produce official statistics.