DOT'S PRE-DISSEMINATION REVIEW PROCESS: A BENCHMARK DATA QUALITY TEMPLATE

The Center for Regulatory Effectiveness
11 Dupont Circle, Suite 700
Washington, DC 20036
www.TheCRE.com

March 2003

DOT'S PRE-DISSEMINATION REVIEW PROCESS: A BENCHMARK DATA QUALITY TEMPLATE

INTRODUCTION: THE IMPORTANCE OF PRE-DISSEMINATION REVIEW

OMB's government-wide data quality guidelines, implementing the Data Quality Act,[1] require that agencies establish a pre-dissemination review process to "substantiate the quality of the information it [the agency] has disseminated...."[2] In discussing the need for the pre-dissemination review process, OMB explains, "Agencies shall treat information quality as integral to every step of an agency's development of information, including creation, collection, maintenance, and dissemination."[3] Thus, the pre-dissemination review process is far more than a simple tick-list of steps that are applied to existing data to determine if it is ready for release; pre-dissemination review is an essential quality assurance process that takes place throughout the development and analysis of information disseminated by an agency.

ABOUT THE CENTER FOR REGULATORY EFFECTIVENESS

The Center for Regulatory Effectiveness ("CRE") is a regulatory watchdog which provides public oversight of and seeks to improve the quality of federal and private sector regulatory activities. To further improve the quality of the federal regulatory process, CRE was the leading proponent of the Data Quality Act.[4] CRE has also participated extensively in the public process of developing government-wide and agency-specific implementing guidelines, including the Department of Transportation's Information Dissemination Quality Guidelines (DOT Guidelines.)[5] As part of its watchdog activities, CRE analyzes and develops recommendations on data quality-related issues. To this end, CRE has analyzed DOT's pre-dissemination review process to ascertain whether its detailed requirements set a benchmark

---

[1] 44 USC 3516 Statutory and Historical Notes.

[2] 67 FR 8459.

[3] Ibid.

[4] *Federal Times*, "Nixon's 'Nerd' Turns Regulations Watchdog," November 11, 2002, p. 22, http://thecre.com/pdf/20021111_fedtimes-tozzi.pdf.

[5] DOT Guidelines, p. 1.

standard that OMB should encourage other agencies to emulate as they refine their own processes. CRE will be reporting on our findings on our website, www.TheCRE.com.

THE DEPARTMENT OF TRANSPORTATION'S PRE-DISSEMINATION REVIEW PROCESS

There are six components to the pre-dissemination review process specified in the DOT guidelines.[6] This pre-dissemination review process is applicable to; the Office of the Secretary (OST), Bureau of Transportation Statistics (BTS); Office of Inspector General (OIG); Federal Aviation Administration (FAA), Federal Highway Administration (FHWA), Federal Motor Carrier Safety Administration (FMCSA), Federal Railroad Administration (FRA), Federal Transit Administration (FTA), Maritime Administration (MARAD), National Highway Traffic Safety Administration (NHTSA), Research and Special Programs Administration (RSPA), Saint Lawrence Seaway Development Corporation (SLSDC), and Transportation Administrative Service Center (TASC). BTS, through their participation in an inter-agency project to improve the quality of statistical information, is also subject to additional statistical standards. BTS explains their additional obligations in a separate document published by the Bureau, "Statistical Policy and Research."

The six pre-dissemination review steps specified by DOT are:

1. Review and consultation;

2. Compliance verification;

3. Maintaining records of additional standards applied to influential information;

4. Ensuring the entire information product fulfills the agency's stated intentions and that the conclusions are consistent with the evidence;

5. Indicating the origin of data; and

6. Ensuring each program office can supply additional data on any subject for which there is a covered information dissemination.

The First Component: Review and Consultation

DOT's pre-dissemination review process requires that agencies allow "adequate" time for review, consistent with the standards required for the type of information to be disseminated.

---

[6] DOT Guidelines, pp. 19-20.

Thus, agencies need to allow more review time for more significant information. This requirement is fully consistent with OMB's government-wide data quality guidelines which state, "The more important the information, the higher the quality standards to which it should be held...."[7]

In addition to requiring adequate time for review, the DOT guidelines impose a positive duty on agencies to consult with stakeholders including, but not limited to, the public, other DOT organizations and State governments[8]. The consultation requirement is important for two reasons:

1. Quality. Consulting with diverse stakeholders through formal and informal processes will enable the agencies to detect errors, biases and other data quality flaws of which they may not have otherwise been aware. The result of the consultations should be an enhancement of overall data quality. Such data quality-related consultations are particularly important for information developed pursuant to rules which were proposed prior to publication of the Department's guidelines as agencies will not have had the opportunity to consult with stakeholders on data quality issues during the rulemaking.

2. Transparency. One of the paramount goals of the Data Quality Act is to enhance the transparency of government processes. Such transparency not only helps detect errors but also and equally important, "is that the public will be able to assess how much of an agency's analytic result hinges on the specific analytic choices made by the agency."[9] Thus, the guidelines should increase the transparency of the federal decision-making process, at least with respect to decisions based on analysis. Although the OMB and DOT guidelines provide a number of other specific requirements to ensure transparency, by consulting with stakeholders about data quality issues, the Department will further enhance and promote transparency.

The Second Component: Compliance Verification

The pre-dissemination review process requires that agencies verify that covered information complies with the DOT guidelines as well as with the other guidance and procedures issued by

---

[7] 67 FR 8452.

[8] DOT Guidelines, p. 19.

[9] 67 FR 8456.

DOT agencies. This compliance verification component is, in many respects, the heart of the pre-dissemination review process since it is an essential step for ensuring that only information that meets data quality requirements is disseminated.

An integral element of the DOT guidelines with which agencies must verify compliance are the detailed principles and guidelines for statistical data located in Appendix A of the document. The Appendix is, "a subset of the DOT Information Dissemination Quality Guidelines."[10] Appendix A, "incorporate[s] the statistical aspects of the OMB guidelines as a baseline and elaborate[s] on its recommendations to produce statistical guidelines adapted for the Department of Transportation."[11] The detailed data quality requirements in Appendix A, "apply to all statistical information that is disseminated on or after 1 October 2002 by agencies of the Department of Transportation..."[12] The Appendix also explicitly states that the guidelines apply to third-party data disseminated by the Department, as part of a covered dissemination, with the exception of the data acquisition requirements if the data was acquired by a non-federal source. The guidelines apply to "reporting systems, surveys, and special studies."[13]

The DOT statistical guidelines are built on three primary components:

1. Structured planning;

2. Sound statistical methods; and

3. Openness.

To ensure clarity, the discussion of each component has three elements:

1. A statement of principles;

2. Guidelines, which are specific recommended actions; and

3. References.

---

[10] DOT Guidelines, p. 1-3.

[11] DOT Guidelines, p. 1-2.

[12] Ibid.

[13] Ibid., p. 1-3.

Clarity is further heightened through the ample use of examples, notes, flow charts and warnings.

Structured Planning

Data System Objectives. The first aspect of planning addressed by the guidelines are Data System Objectives. The principles articulated by DOT highlight that development of the data systems will be driven by the sponsoring organization's strategic plan and budget. Other factors influencing system development are the data users and the primary questions to be answered by the system. Of particular note, DOT explains that objectives of the system should "describe what federal programs and external users will accomplish with the information."[14] The guidelines go on to explain that these expected accomplishments should be traceable to goals in the sponsoring organization's strategic plan. Thus, the principles provide guideposts for ensuring that the planned data system will have "utility" as defined by the OMB guidelines, i.e. that the data is useful to the intended users. Data systems that are not able to demonstrate that they accomplish clear goals tied the agency's strategic plan would lack utility and would not comply with data quality guidelines even if the numbers generated were, in an abstract sense, accurate.

DOT provides a series of guidelines to carry out its stated principles for data systems. The guidelines are intended to ensure that the data system is developed in conjunction with users and other stakeholders so as to meet their specific needs. Specifically, the guidelines state that the responsible agency "should develop and update the data system objectives in partnership with critical users and other stakeholders." The guidelines go on to state the objectives of the data system should, "indicate each major need that will be fulfilled by the system and the data users associated with that need, and the key questions that will be answered by the data."[15] Thus, the DOT guidelines call for a lean systems designed to meet specific users needs and answer specific questions. These guidelines clearly aid in helping to ensure that data systems fulfil the Act's requirement that disseminated information possess utility.

Data Requirements. The principles in DOT's data quality guidelines emphasize that decisions regarding data collections, including the data elements and collection methodologies, should be tied to specific "measurement concepts." Measurement concepts, defined as a characteristic of people, objects, businesses or events, thus drive the determination of data requirements. According to the principles, data requirements are created in order to measure each measurement concept. The principles also point out that, "in addition to data that are directly related to strategic plans, additional data may be required for possible cause and effect

---

[14] Ibid., 1-4.

[15] Ibid., 1-5.

analysis."[16] As an example of this principle, the document cites the collection of weather data for causal analysis of traffic accidents. The collection of additional explanatory data is in keeping with the OMB guidelines' discussion of objectivity which note that, in some instances, "other information must also be disseminated in order to ensure an accurate, clear, complete, and unbiased presentation."[17] The DOT data quality principles demonstrate that objectivity is achieved by designing the prerequisites for objectivity, such as causality-related information, into the data system during its earliest planning stages.

The detailed guidelines and examples that DOT provides for implementing the Data Requirements principles emphasize data practices that are important for compliance not only with the data quality guidelines but also with for compliance with other "Good Government" laws. For example, the guidelines state that, although all reasonable measures should be initially considered when determining how to quantify a measurement concept, the final data choice decision "will be made based on ease of acquisition, constraining factors (e.g. cost, time, legal factors), and accuracy of available data."[18] By recognizing that cost is a factor that needs to be considered when determining data collection requirements, the data quality guidelines also embody key principles in the Paperwork Reduction Act and the Executive Order on regulatory review. The DOT guidelines also encourage standardization with other databases and use of coding standards. These guidelines, therefore, also help support compliance with the National Technology Transfer and Advancement Act and OMB Circular A-119.

Data Acquisition Methodology. In its discussion of data acquisition methods, DOT reiterates the need to consider cost, accuracy and time restrictions. DOT goes on to discuss the need for agencies to obtain appropriate expertise, particularly for complex data systems. As the guidelines explain, "As the process gets more complex, there is no substitute for expertise."[19] Should the requisite expertise not be available in-house, the guidelines recommend either consulting with another agency or using a contractor.

Data Sources. DOT's principles with regard to the sources of data encourage economic efficiency by recommending the use of existing data sources, where possible. Furthermore, the guidelines also recommend determining if existing data collections can be modified to

---

[16] Ibid., 1-7.

[17] 67 FR 8459.

[18] DOT Guidelines, p. 1-8.

[19] Ibid., p. 1-10.

meet additional needs. The use of third-party data is also suggested as a possible source of information.[20] By stressing the need to be efficient and avoid redundant data collections, the DOT guidelines provide valuable guidance in helping agencies comply with the Paperwork Reduction Act.

The DOT data source guidelines provide practical advice on how to implement the previously elucidated principles. For example, in order to avoid needless new data collections, the guidelines encourage agencies to, "Research whether government and private data gathering systems already have data that meets your needs."[21] Thus, guidance that would enhance agency compliance with a key theme of the Paperwork Reduction Act, avoiding redundant information collections, is woven throughout this section of the DOT guidelines.

Data Collection Design. DOT's principles for designing data collections highlight that the design process is "one of the most critical phases in developing a data system." The principles go on to highlight the fact that the accuracy of the data, and by extension, the estimate derived from the data, "are heavily dependent on the design of data collection."[22] The guidelines thus call attention to a crucial yet often overlooked element of the overall process of developing a data system.

The guidelines implementing the Department's design principles provide important advice on data sampling. The guidelines include a prominent warning about sampling techniques that might yield biased data. Thus, these guidelines again demonstrate how achieving the specific objectives of the Data Quality Act, such as objectivity, requires that data quality principles be applied to data development from the earliest stages – long before it is disseminated to the public.

Sound Statistical Methods

After the data collection has been designed, the next step in the process of developing a data system is to actually collect the data. DOT points out that "The physical details of carrying out the collection are critical to making the collection design a reality."[23]

---

[20] Ibid., pp. 1-11:1-12.

[21] Ibid., p. 1-12.

[22] Ibid., p. 1-13.

[23] Ibid., p. 1-16.

Data Collection Operations. Minimizing error is the focus of DOT's principles concerning the collection of data. DOT's principles highlight potential sources of error ranging from deviations in collection from design to errors from third-party collection of data to errors resulting from converting paper records to electronic formats.[24] By creating awareness of potential sources of error and providing precepts for controlling such errors, the principles provide DOT agencies with practical standards that need to be incorporated into their data collection exercises. Failure to follow the principles could well result in significantly flaws being introduced during the data collection process which, unless corrected, could corrupt analyses of the data.

The guidelines for implementing the data collection principles focus on practices to minimize the potential error sources identified in the principles, such as use of verification systems when paper data is converted to electronic format. The guidelines also contain important advice for minimizing missing and erroneous data such as proper training of interviewers and other data collectors, making the data collection process as easy as possible for data collectors, and ensuring that forms and file layouts are well designed.[25] By focusing on the human factors that affect data quality, the guidelines provide advice that is solidly practical rather than simply being theoretically sound.

Avoiding Missing Data. The DOT guidelines recognize that missing data is inevitable and explain the two main types of missing data; item-level and unit-level. Item-level missing data refers to missing data elements in a report while unit-level missing data refers to a report that is missing. The principles are designed to minimize the occurrence of missing data. One of the guidelines' key principles is to perform follow-up work which can "dramatically reduce the incident of both unit-level and item-level missing data."[26]

In addition to providing recommendations on how to minimize missing data, the guidelines provide guidance on how to manage the issue of missing data. For example, the guidelines direct agencies to document and clearly post with the data, or disseminated output from the data, the missing data avoidance procedures that were employed.[27] Providing the public with information about missing data avoidance procedures used in a data collection project is important for transparency since it will allow the public to evaluate whether those procedures

---

[24] Ibid., pp. 1-16:1-17.

[25] Ibid., pp. 1-17:1-18.

[26] Ibid., p. 1-19.

[27] Ibid.

were sufficient or whether they could have introduced any unintended consequences, such as introducing bias, into the database and subsequent analyses.

Data Editing and Coding. The next step after data collection is data processing. The first elements in the processing function are editing and coding the data. The principles explain that the purpose of the editing function is to "identify missing , invalid, duplicate, inconsistent entries, or otherwise point to data records that are potentially in error."[28] The DOT principles explain that although outlying data points should be examined for possible correction, they are not necessarily bad data and should not be automatically deleted. The guidelines provide additional details about both data editing and coding as well as illustrative examples to guide agencies in developing their editing and coding process. Although editing is described as a "final inspection-correction method" and note that it "is almost always necessary" the principles stress that "data quality is better achieved much earlier in the process..."[29] The principles make clear that although data editing and coding is an important element in ensuring overall data quality, it is no substitute for using proper data quality practices throughout every stage of the information process.

The guidelines emphasize that "every data collection" as well as all third-party data, should be subjected to a data editing process. Also recommended is that consideration be given to use of "statistical edits...to detect more subtle errors."[30] In addition to various recommendations on the editing and coding process, the guidelines note that "the editing and coding process should clearly identify missing values" since use of zeros or blanks for missing values "have historically caused confusion."[31] The guidelines also state that, "The editing and coding process and editing statistics should be documented and clearly posted with the data, or with the disseminated output of the data."[32] Thus, as was the case with missing data procedures, the guidelines demonstrate that transparency requires that public be informed how agency and third party data is handled at every stage of the process.

Handling Missing Data. Preventing bias and other forms of error is the focus of the DOT principles for handling missing data. The principles for handling missing data inform agencies that "missing data can introduce serious error into estimates." Furthermore, the principles

---

[28] Ibid., p. 1-21.

[29] Ibid., 1-22.

[30] Ibid., 1-23.

[31] Ibid.

[32] Ibid., p. 1-24.

explain that often "there is a correlation between the characteristics of those missing...resulting is biased estimates." For this reason, corrections need to be made to "mitigate his damage."[33] The principles outline techniques, such as imputation, for addressing missing data.

The missing data guidelines provide advice on statistical procedures for handling missing data under various circumstances. References are provided for additional guidance, including Statistical Policy Working Paper from OMB's Office of Information and Regulatory Affairs. Two key issues that are discussed in the guidelines are:

1. Documentation. The guidelines explain that the methodology for addressing missing data, "should be fully documented and summarized in the data systems source and accuracy statement."[34] The guidelines also call for identifying the data that was imputed so as to help evaluate the impact of manipulations and that this information should also be included in the source and accuracy statement.

2. Analysis of impact of missing data. The DOT guidelines state that the effect of missing data should be analyzed. Included in the detailed DOT guidelines is a discussion of the minimum standards for the analysis of missing data as well as guidance on how frequently such analyses should take place.[35]

The DOT guidelines for missing data advance two essential aspects of data quality; minimizing bias and other errors, and ensuring transparency so that users of the data, or reports based on the data, are able to make their own independent evaluation as to quality of the disseminated information.

Producing Estimates and Projections. The principles explain the terms estimate, "an approximation of some characteristic of the target group..." and projection, "a prediction of outcome from the target group..."[36] The guidelines also discuss derived data, data calculated directly from the information collection, added from a separate source or a combination of the two. Of particular note, the principles explain that derived data, "is a way to enhance the

---

[33] Ibid.

[34] Ibid., p. 1-25.

[35] Ibid., pp. 1-25:1-26.

[36] Ibid., p. 1-26.

data set without increasing respondent burden or significantly raising costs."[37] By highlighting the potential for minimizing the burden and cost of information collections while improving their utility, the principles demonstrate how the DOT data quality guidelines can be used to enhance compliance with the Paperwork Reduction Act and, thus, the concordance between the PRA and the Data Quality Act.

The theme of using derived data to enhance the information collection without additional burden is reinforced by the implementing guidelines. The guidelines also discuss other statistical practices including the importance of accompanying any estimates from sample with the standard error. The guidelines explain that the calculation of the standard error needs to take into account sample design for the collection. When more complex sample designs are used, the guidelines direct agencies to, "use replicated methods (e.g., jackknife, successive differences) incorporating sample weights. Consult with a variance expert."[38] Similarly, the guidelines direct that any statistical software used in developing estimates and their standard errors utilize techniques that take the design of the data collection into account. The guidelines thus provide detailed directions for carrying out the OMB requirement to use sound statistical practices. In keeping with transparency requirements, the guidelines also state that all estimation and projection methodologies should be documented and provided with the disseminated information.[39]

Data Analysis and Interpretation. DOT's principles for analyzing and interpreting data cover a number of statistical issues ranging from probability theory to the circumstances for using time series analytic methodologies to when to use spacial data analysis techniques. A special note is made that analytic methodologies such a linear regression "assume independence of the data points which may them invalid in time and geographic cases."[40] The principles also define "robustness" and note that it is "a critical factor in planning an interpreting analyses."[41] One of the key principles annunciated is that the "stability of the process being analyzed" should be taken into account when interpreting data. The principles explain that if the process is not stable, i.e. if it has significantly altered since the data was collected, analytic results may be of limited value for decision-making purposes. Thus, even when a data collection and

---

[37] Ibid.

[38] Ibid., p. 1-27.

[39] Ibid., p. 1-28.

[40] Ibid., p. 1-29.

[41] Ibid.

analysis is otherwise sound, depending on the underlying process being measured, the analytic results may have limited or no utility.

The DOT data quality guidelines provide a series of precepts for implementing the principles for data analysis and interpretation:

1. Project plan. A project plan for analyzing data should be developed for all but the most simple analyses. Even "exploratory analyses" should be conducted according to an analytic plan.

    - The plan should be reviewed by subject matter experts "to ensure that the analysis is relevant to the questions that need answering."[42]

    - The plan should also be reviewed by data analysis experts, even if the plan was written by such an expert, "to ensure proper methods are used."[43]

2. Appropriate and well documented statistical methodology. "All statistical methods used should be justifiable by statistical derivation or reference to statistical literature."[44]

    - The analytic process "should be accompanied by a diagnostic evaluation of the analytic assumptions."[45]

    - A study of the probability that there will be violations, of various degrees, of statistical assumptions – and the impact such violations would have on the conclusions – should also be included.[46]

---

[42] Ibid.

[43] Ibid.

[44] Ibid., p. 1-30.

[45] Ibid.

[46] Ibid.

- There should be thorough documentation in the project plan and the final report of all "statistical methods, derivations or references, assumption diagnostics, and robustness checks..."[47]

The guidelines also provide a partial list of the analytic techniques available to agencies.

3. Sample design. When a complex sample design is used to collect data, that design should be incorporated into the analytic methods, "via weights and changes to variance estimation (e.g., replication)."[48]

4. Additional variables. In order to assist in interpreting the relationship between two or more variables, related variables should be included in the analysis. The guidelines explain that "[m]issing important variables can lead to bias. A subject matter expert should choose the variables."[49]

5. Documentation. Analytic results should be documented and included, or concurrently posted, with any report using those results. The report should be focused on the questions answered by the data analysis, the methodologies and assumptions used and the limitations of the analysis.[50]

- A statement on the limitations of the analysis "should always" be included.
- The analysis should be worded to "reflect the fact that statistically significant results are only an indication that the null hypothesis may not hold true. It is not absolute proof." Likewise, the wording should indicate that, if a test "does not show significance, it does not meant hat the null hypothesis is true, it only means that there was insufficient evidence to reject it."[51]

---

[47] Ibid

[48] Ibid.

[49] Ibid.

[50] Ibid.

[51] Ibid.

6. Confidence intervals. Confidence intervals or other tests that are based on a sampling concept should not be applied to analytic results based on 100% data. Such tests should, instead, be used to measure "the variability of the underlying random phenomenon."[52]

7. Stability. The stability of the process being analyzed should be considered and commented on when interpreting the analysis. Thus, if a process, such as airport security, has significantly changed since the analysis was conducted, the results of the analysis may be of be of very limited value.[53]

Openness

Information Dissemination

DOT explains that there are two key points with regard to the dissemination of statistical information.

1. Openness. Because openness "is relative to all aspects of data quality...statistical information being disseminated must be accompanied by documentation."[54]

2. Final Pre-dissemination review. The final quality reviews of information before dissemination "are final assurances that all quality control steps have been taken and that the dissemination package is complete."[55] It is important to note that by discussing the final quality reviews, the guidelines implicitly highlight the importance of quality reviews throughout the study design, implementation and analysis processes.

Publications and Disseminated Summaries of Data. DOT's principles are built on the principle of ensuring both transparency and utility of disseminated information. Thus, the principles call for information "to be clearly presented to users" and for users to be "informed about the

---

[52] Ibid., pp. 1-30:1-31.

[53] Ibid., p. 1-31.

[54] Ibid., p. 1-32.

[55] Ibid.

sources of the information presented."[56] To prevent confusion or misinterpretation, the principles also call, to the extent practical, for tables, graphs and other visual aids to be useful "as stand-alone products in case they become separated from their original context" and for users to be provided with the methodology used to produce the tables and other materials. An additional important transparency principle is that data users be informed about the degree of uncertainty of statistical interpretations.[57]

Clarity of communication is a key goal of the DOT guidelines implementing the principles associated with publications and the dissemination of data summaries. To achieve this goal, the guidelines provide specific recommendations, based on the GPO Style Manual, as to basic, if often overlooked, measures such as ensuring that:

- Documents are well organized and well written;

- Tables, graphs and other visual aids are consistent with each other and with the associated text;

- Clearly worded titles are used for tables, graphs and other visual aids. The titles should answer three questions: what data is being presented; the geographic area being presented and the time frame covered;

- Detailed source reference information should be provided for information in tables, graphs and other visual aids as well as for text that provides data which is not in the illustrative materials. Since information sources, such as databases, may be updated, the source information should include an "as of" date;

- Footnotes should be used as appropriate;

- Information disseminated via the internet should meet applicable accessability requirements;

- Documents should include or reference the methodology information discussed in the sections of the guidelines concerning the development of estimates and projections and data analysis and interpretation; and

---

[56] Ibid.

[57] Ibid.

- Specific contact information should be provided in order to facilitate comments and recommendations from users of the information.[58]

The recommendation regarding contact information is particularly significant since it demonstrates that communications flows are a two-way process and that obtaining stakeholder input is an essential part of the communications process.

Micro Data Releases. Helping the public determine whether results are reproducible is one the key purposes of releasing micro data.[59] The DOT guidelines describe the term "micro data" as referring to "data files with various information at the 'unit' level."[60] Examples of micro data include individual responses to a survey and reports of individual incidents. Although release of micro data is supported to assess reproducibility and enhance the usefulness of information disseminations, micro data may not be released if it would violate confidentiality. The DOT guidelines state, "micro data should not be released in violation of existing protections of privacy, proprietary information, or confidentiality."[61]

DOT has a series of guidelines providing practical advice on how to implement the Department's principles for micro data releases. The guidelines include ensuring that micro data releases are compatible with a range of commonly available software and providing technical information, such as the file layout and related information so as to ensure that the data is useable. The guidelines call for micro data releases to be accompanied by or provide reference to the "quality-related documentation discussed herein: planning documentation and collection, processing, and analysis methodology."[62] The guidelines also explain that readers should be provided with a point of contact for comments and questions. Thus, the guidelines are consistent in their call for transparency and two-way communications between agencies disseminating information and the users of that information.

Source and Accuracy Statements. To provide the public with information about the quality of data being disseminated, the DOT has established principles calling for Source and Accuracy Statements (S&As). The S&As are "compilations of data quality information..." that "provide

---

[58] Ibid., pp. 1-32:1-33.

[59] Ibid., p. 1-33.

[60] Ibid.

[61] Ibid., p. 1-34.

[62] Ibid.

information on where the data came from, how it was collected, and how it was processed."[63] An additional key principles of the S&As is that they "include information on known strengths and weaknesses of the data."[64] S&As should be regularly updated to account for any changes in methodology and "results of any quality assessment studies." Thus, DOT has made providing the public with current and informative assessments of the quality of data being disseminated a data quality principle.

DOT has provided a series of guidelines for implementing the principle of informing stakeholders about the quality of the data being disseminated. These guidelines detail the information to be included in the S&As, such as data collection design and major sources of error.[65] DOT's policy statements regarding S&As make clear that data quality goes beyond making sure that disseminated information meets certain minimum standards to ensuring that the public is informed about the quality, including the strengths and weaknesses, of the information that is disseminated.

Pre-Dissemination Review. DOT provides three basic principles for pre-dissemination reviews:

1. Use of formal and informal reviews. Use of formal and informal reviews for publications, summaries and micro data "will help ensure that a data product meets a minimal level of quality."[66]
2. Diverse reviewers. In that there are multiple aspects of quality in a final information product, "reviews need to be conducted by several people with different backgrounds."[67]

3. Reviews throughout development. Rather than simply reviewing the final information product, reviewing documentation created "through the various stages of data development will enhance the review process."[68]

---

[63] Ibid.

[64] Ibid.

[65] Ibid., 1-35.

[66] Ibid,

[67] Ibid.

[68] Ibid.

DOT provides a series of detailed guidelines for implementing the Department's pre-dissemination review principles which require the participation of various specialists, some of which need to be otherwise independent of the data collection and analysis exercise. The DOT guidelines, by covering a variety of procedures and precautions, demonstrate a thorough understanding of the diverse facets of data quality. Specific pre-dissemination review guidelines include:

1. Independent Subject Matter Specialist. A specialist in the relevant subject matter, who is not directly involved in the information collection and analysis, "should review the plans, methodology documents, and reports prior to publication. They should also review publications and summaries resulting from the data for content and consistency."[69]

2. Independent Statistical Expert. A data analysis expert, such as a statistician, who was not directly involved in the data collection and analysis should review the planning and methodology documents and reports prior to dissemination. The expert should also "review publications and summaries resulting from the data for the wording of statistical interpretation."[70]

3. Information Technology Specialist. With regard to micro data, "the release files and the metadata should be reviewed by an information technology specialist for clarity and completeness."[71]

4. Visual Information Specialist. A style and visual information specialist should review publications "for compliance with style standards."

5. Accessability Specialist. Information disseminated through the internet should be reviewed for accessability by a specialist in compliance with Section 508 of the Rehabilitation Act

---

[69] Ibid.,1-36.

[70] Ibid.

[71] Ibid.

6. Accessability and Interpretability Testing. Data which is disseminated over the internet using special software, "should be tested for accessability and interpretability prior to dissemination."[72]

7. Peer Review Qualifications. When an external peer review is used:

   - Peer reviews should be "conducted in an open and rigorous manner";
   - Selection of the peer reviewers should be primarily based on "necessary technical expertise."; and
   - Persons engaged in the peer review "should be expected to disclose to DOT prior technical/policy positions they may have taken on the issues at hand and their sources of personal and institutional funding (private or public)".[73]

By setting standards for peer review processes, including the requirement that the peer review process be open and rigorous, DOT has implemented OMB's requirements for peer review transparency.[74] The Department's policy of using statistical, subject matter and other experts to review virtually every aspect of the data collection, analysis and dissemination process is a further demonstration of DOT's commitment to implement OMB's precept that information quality is integral to every step of the information process.

Evaluating Information Quality

Regular assessments of all aspects of data quality are needed in order to attain and maintain "a high level of information quality." Thus, to ensure high quality information, the DOT guidelines call for "regular assessments of the data collected, special studies of the data and the effectiveness of the collection and processing processes, and quality control of key processes to both control the quality during operation and to collect data quality information."[75] DOT has developed a series of principles and implementing guidelines for evaluating data quality on an ongoing basis.

---

[72] Ibid.

[73] Ibid.

[74] 67 FR 8454.

[75] DOT Guidelines, p. 1-37.

Assessing/Auditing Data Quality. Data Quality Assessments are defined by DOT as "data quality audits of data systems and the data collection process."[76] With a goal of helping to improve data quality, these comprehensive assessments are intended to assess the extent to which an information system is adhering to the DOT guidelines and to evaluate error sources as well as other potential quality problems with the data. A key aspect of the assessments is that they conclude with recommendations for improvements.[77]

The implementing guidelines for data quality assessments provide roles for both the sponsors and the users of the data system. Initial data quality assessments are to be made by the project's sponsor since they have the greatest access to and information about the data system. These assessments are to be undertaken periodically to ensure that disseminated information meets requirements. The assessment should be used in any redesign of the data system. The guidelines also call for the assessment team to include at least one individual who understands data quality but who was not involved in developing the data system. Users of the system, including secondary users, are to be consulted on the assessment. Specifically, the data users should provide:

1. Recommendations of areas to be assessed; and

2. Feedback on the utility of the data products.[78]

DOT also calls for documentation of the findings and results of the assessment.[79]

Evaluation Studies. The DOT guidelines recognize that "many aspects of data quality cannot be assessed by examining end-product data."[80] It is for this reason that the guidelines call for conducting evaluation studies which are "focused experiments" to assess some element of data quality. DOT's principles provide a series of examples of evaluation studies including: user surveys, re-measurement, independent data collection, collection method parallel trials (incentive tests), census matching, administrative record matching, comparisons to other data

---

[76] Ibid.

[77] Ibid.

[78] Ibid.

[79] Ibid., p. 1-38.

[80] Ibid.

collections, methodology testing in a cognitive lab and mode studies.[81] The principles also explain that "critical data systems" are those systems containing "influential" information or that provide data to Department-level performance measures.

DOT guidelines provide two key recommendations specific to critical data systems:

1. A program of periodic evaluation studies to "estimate the extent of each aspect of non-sampling error..."

2. Periodic evaluations of:

   - Bias resulting from missing data
   - Coverage bias;
   - Measurement error; and
   - User satisfaction.[82]

Thus, DOT is implementing the concept developed in the OMB guidelines that more important data requires higher data quality standards. In addition to recommendations specific to critical data systems, DOT also has implementing guidelines for evaluating all data systems including conducting evaluation studies when:

1. There is an indication that data elements may not be meeting data quality standards due to one or more sources of error; and

2. An analysis of the data demonstrates a problem even though the cause of that problem is not obvious.[83]

Quality Control Systems. Since each data collection and processing activity will add some amount of error, each activity needs some type of quality control system to prevent or correct the error. Two factors which will influence how elaborate the quality control system needs to be are: the importance of the data being collected/processed; and how complex and/or

---

[81] Ibid.

[82] Ibid.

[83] Ibid.

tedious the collection/processing activity with more complex/tedious tasks requiring more controls. Examples of quality control systems provided by DOT include: 100% replication (such as of key entry); sample replication which can be used in a stable ongoing process; analysis of the data files both before and after the activity in question; and simple reviews.[84]

DOT's guidelines for implementing quality control systems begin with requirement that every activity "should be examined for its potential to introduce error."[85] The extent of quality control should be based on the two principles that greater control is placed on data that is more important and/or has a greater the likelihood of error. The guidelines also specify that data from quality control systems should indicate the effectiveness of the effort. As with virtually every single data related activity, the guidelines call for documentation; quality control should be included in the methodology documentation.

Correcting Data Errors. DOT starts with the principle that "no data system is free of errors."[86] The actions taken to correct errors are dependent on:

1. The strength of the evidence that there are one or more errors;

2. The impact "a potential error would have on the primary estimates" generated by the data system; and

3. The resources that would be required to verify and correct the error(s).[87]

A standard process for dealing with errors "should exist and be documented."[88] If known errors cannot be corrected, such as when a data file is "frozen," the errors should be documented and should accompany the flawed data.

The Third Component: Documentation of Influential Information

---

[84] Ibid., p. 1-39.

[85] Ibid.

[86] Ibid., p. 1-40.

[87] Ibid.

[88] Ibid.

The DOT guidelines require that agencies maintain records of the additional standards are applies to information considered to be influential.[89] Thus, there are two key issues with respect to influential information:

1. The Department's definition of information considered to be influential; and

2. The additional quality standards applied to the information.

Definition of Influential. DOT, in keeping with the OMB guidelines, sets specific attributes defining the term influential. The key requirement is that the information have a "clear and substantial impact" on important private sector decisions or public sector policies.[90] DOT explains that this means, among other requirements, that the expected impact must be highly likely to occur rather than speculative. DOT also notes that, based on OMB's definition, the term influential only applies to scientific, financial or statistical information. Thus, a policy decision, no matter how important, would not be considered as influential information although at least some of he information in the administrative record underlying that decision might potentially be classified as influential.

In regulatory contexts, the Department relies significantly on the Executive Order on regulatory review for assessing whether information is influential. Specifically, DOT states that influential information is "information that can reasonably be regarded as being one of the major factors in the resolution of one or more key issues in a significant rulemaking, as that term is defined in Executive Order 12866."[91] The Executive Order defines a "significant regulatory action" as one which is likely to have at least one of the four impacts[92]:

1. Economic. An annual impact on the economy of at least $100 million or a material adverse affect on the economy, a sector of the economy, productivity, jobs, competition, public health or safety, the environment or state, local or tribal governments or communities.

2. Interagency conflict. A serious inconsistency or otherwise interfere with an action already taken or planned by another agency.

---

[89] Ibid., p. 19.

[90] Ibid.. [emphasis in original].

[91] Ibid.

[92] 58 FR 51738.

3. Budgetary Impact. Materially altering the budget impact of entitlements, grants, user fees, or loan programs or affect the rights and obligations of the recipients of such programs.

4. Novel Issues. Raise novel legal or policy issues rising out of legal mandates, the President's priorities or the principles in Executive Order 12866.

In non-regulatory contexts, DOT with consider two aspects of information when considering whether information is influential, the breadth of the expected impact and the intensity of expected impact[93]. Thus, information which affects a broad range of parties, such as an industry, is far more likely to be considered as influential than information affecting a narrow range of parties, such as a given company, even if the impact on that company is significant. With regard to the intensity of impact, DOT will be guided by the same types of factors that determine whether a rulemaking is significant. However, these factors are considered as guidance and actual decisions will be made on a case-by-case basis.

DOT explains that certain classes of information may be defined as influential. Decisions regarding other information will be made on a case-by-case basis. DOT agencies are encouraged to apply the influential label "only when clearly appropriate."[94]

Standards for Influential Information. DOT has adopted the quality principles standards set forth in the Safe Drinking Water Act amendments of 1996 for analyses of risks to human health, safety and the environment[95]. This use of the SDWA standard is in addition to the specific principles and guidelines relevant to influential information, such as the requirements for "critical data systems" that are contained in the DOT's detailed statistical guidelines and were discussed in the Compliance Verification section above. The SDWA requirements adopted by DOT include:

1. Use of the best available science and supporting studies; and

2. Use of data collected by the accepted or best available methods;

Furthermore, when disseminating influential risk-related information, the SDWA language requires that the information include:

---

[93] DOT Guidelines, p. 27.

[94] Ibid., p. 28.

[95] Ibid., p. 16.

1. Each population addressed by any estimate of applicable effects;

2. The expected or central estimate of risk for the specific affected populations;

3. The upper and lower bounds of the risk estimate;

4. Each significant uncertainty identified in the risk assessment and studies that would help reduce the uncertainty; and

5. A list of studies known to the agency which support, are directly relevant to or fail to support the findings of the assessment and the methodology used to reconcile inconsistencies in the scientific data.

With regard to the reproducibility of information, the DOT guidelines state that they intend to follow a policy determining, in consultation with relevant scientific and technical communities, when reproducibility standards should be applied to original and supporting data.[96] Furthermore, with regard to analytic results, DOT's policy is to "favor sufficient transparency about methods to allow independent reanalysis by qualified members of the public."[97] When confidentiality requirements obviate such transparency, DOT's policy is to "apply and document especially rigorous robustness checks."[98] Thus, DOT's policies regarding reproducibility are fully consistent with OMB's government-wide guidelines.

The Fourth Component: Ensuring Information Fulfils Agency Intentions

DOT's pre-dissemination review process includes ensuring, "that the entire information product fulfills the intentions stated and that the conclusions are consistent with the evidence."[99] This component of the pre-dissemination review process can be considered as an opportunity for the agency to look at the "big picture." DOT shows considerable foresight and common sense in understanding that it is possible to lack overall perspective when going through the minutia of a detailed process. Thus, even after the responsible agency has gone through the detailed compliance verification process, they still need to step back and take a big-picture perspective to ensure that an information product makes sense, *i.e.*, conclusions

---

[96] Ibid., p. 15.

[97] Ibid.

[98] Ibid, pp. 15-16.

[99] Ibid., p. 19.

are consistent with the evidence, and that the project's goals were achieved. The detailed statistical guidelines provide important tools, including the extensive documentation and planning requirements, that can be used in ensuring that the final work product fulfills agency intentions. Although DOT does not provide specific guidance on how to ensure such a forest-not-trees perspective, that may be most appropriate for this purpose.

The Fifth Component: Origin of the Data

Knowing the origin of data is one of the basic requirements for transparency. Thus, one of the specific pre-dissemination review steps is to ensure that, when using data from an external source, the origin of the data is indicated.[100] With regard to statistical information, compliance with the comprehensive documentation requirements discussed in the Compliance Verification section of this document should ensure the agency complies with the origin of data requirement. For non-statistical information, this component of the pre-dissemination review process provides the opportunity to ensure that the origin of all third-party data is clearly provided.

The Sixth Component: Having Additional Data

The final component in DOT's pre-dissemination review process is ensuring that "each program office can provide additional data on the subject matter of any covered information it disseminates."[101] This step of the pre-dissemination process serves as a capstone of the overall data quality assurance process. By being able to provide additional data on all covered information disseminations, the Department is demonstrating its intention to be ready to respond to requests from the public and other stakeholders for clarification or additional insight into an information product. Thus, the DOT guidelines recognize that communications are a two-way process and that data quality is one element to assist in the ongoing dialog between the Department and the stakeholders it serves.

CONCLUSIONS

[100] Ibid., p. 20.

[101] Ibid.

1. DOT's pre-dissemination review process is a thorough data quality assurance program fully consistent with the requirements set forth by the Office of Management and Budget.

2. DOT's data quality guidelines provide a comprehensive and elegant exposition on the planning, collection, analysis and dissemination of statistical information.

3. DOT's data quality guidelines recognize that information disseminations are part of a two-way communication between the Department and its stakeholders. In recognition of the importance of this dialog, consultation by DOT agencies with interested parties on the data quality issues is part of the Department's pre-dissemination review process.

4. DOT, through its pre-dissemination review process, has established a benchmark standard for data quality assurance. OMB should strongly consider using the DOT guidelines, including the integral sections on statistical information, as a template for evaluating the data quality programs of other agencies.