Quality Assessment Tool

for Administrative Data

Paul S. Marck[[[1]](#footnote-1)]

U.S. Census Bureau, Research and Methodology Directorate, Quality Program Staff

Abstract: The United States Federal Committee on Statistical Methodology (FCSM) commissioned a group to develop a tool for assessing the quality of administrative data.

The basic motivation for the tool is to help a statistical agency understand the quality of the data they acquire. This will help program agencies appropriately use the data and produce sufficient documentation to comply with Federal quality standards.

To develop the tool, the group researched the practices of a number of international statistical agencies including The Australia Bureau of Statistics, The U.S. Census Bureau, Eurostat, The Office of National Statistics in the United Kingdom, and Statistics Canada.

The group used the most common quality dimensions as a framework for assessing quality. These dimensions included Relevance, Accessibility, Interpretability, Coherence, Accuracy, and Institutional Environment. The group further organized the tool into three phases of data acquisition—discovery, initial acquisition, and repeated acquisition to help manage response burden.

Based on feedback from initial testing with Federal, State, and Third party organizations the group revised the assessment questions and incorporated a new phase for documenting information requirements. The requirements provide a more objective method to assess the data’s fitness for use and help source agencies understand the data needs.

1. Introduction

Statistical agencies publish data to help inform users about various topics of interest. Traditionally, surveys collect data directly from a business or individual. The data collection costs are typically a substantial portion of the expense for conducting a survey or complete enumeration. To help reduce costs and lessen respondent burden a number of mandates have been issued to improve the utility of existing data sources.

In September 2006, the U.S. Office of Management and Budget (OMB) issued Standards and Guidelines for Statistical Surveys.[[[2]](#footnote-2)] Standard 1.1 requires that Statistical Agencies seek out alternative sources of data when planning an information collection. In July 2011, an Interagency Council on Statistical Policy for the U.S. Government requested the Subcommittee for the Statistical uses of Administrative Records to examine the quality of administrative data gathered by government agencies. The results of this effort were issued on February 14, 2014 in an executive memorandum on “Guidance for Providing and Using Administrative Data for Statistical Purposes”[[[3]](#footnote-3)]

Use of alternative data sources however gives rise to concerns about the quality of the data and their fitness for use. To ensure an alternative data source is fit for use the statistical agency must be able to translate the expectations of their users into requirements for their data suppliers.

2. Producer Quality

Statistical agencies are likely to have numerous users, multiple information products, and possibly complex inter-relationships between given products and users that make understanding let alone ensuring quality a challenge. To facilitate the understanding of this process it is helpful to document user expectations and the desired attributes for an information product.

One method a number of statistical agencies use to describe quality is to identify various dimensions of importance. By being cognizant of user needs for the various dimensions, a statistical agency can make a more conscious effort to help ensure the quality of their information products.

For this tool, we use the following dimensions to describe quality: Relevance, Accessibility, Interpretability, Coherence, Accuracy, and Institutional Environment. These dimensions are conceptually similar throughout various statistical agencies such as U.S. Census Bureau [[[4]](#footnote-4)], Statistics Canada [[[5]](#footnote-5)], Australian Bureau of Statistics [[[6]](#footnote-6)], The United Kingdom’s Office of National Statistics [[[7]](#footnote-7)] and other Eurostat [[[8]](#footnote-8)][[[9]](#footnote-9)] agencies. There are some differences primarily in how the dimensions are grouped. For instance, this tool combines timeliness with relevance where most other agencies treat timeliness as a separate dimension. Similarly, this tool separates interpretability from accessibility where Eurostat agencies group accessibility and clarity.

3. Supplier Quality

Traditionally, statistical agencies collect data directly from persons and companies. As the data collection process shifts to using administrative data suppliers, the statistical agency now becomes the customer. The various user needs and attributes identified for an information product now become requirements for data suppliers. The role of the statistical agency now changes to assessing if an alternative data source is fit for use.

The most likely alternative source is another agency within the government. Data from these agencies are typically ideal for a number of reasons. Agencies are likely to share common leadership that can direct or at least facilitate data sharing arrangements and promote the use of good quality practices. A single agency may also have data for a complete population of interest. The data can also be very accurate, particularly when directly related to the agency’s purpose.

Another alternative source comes from commercial records, such as data from a utility company about its customers. This data can be quite accurate depending on how relevant the data is to the business needs of that organization. The number of organizations to contact to get the same coverage for the previous options however may be greater but these data sources can still offer efficiencies in data collection. A few limitations include possible financial expenses to acquire the data and limited insight into the data collection and processing methods of the organization.

One last alternative source comes from a host of social media or other compilation sites. Data quality issues blossom here. Accuracy of source information may be unknown. The number of data sources is greater than the previous options and they are likely to have biases. There is also the potential for definitional differences and gaining any oversight into the process is unlikely.

When collectively reviewing the various data suppliers a few common observations can be made. First, when a supplier has data of interest to a statistical agency, it is likely that the supplier may also share an interest in the information product of the statistical agency. This shared interest can be instrumental to building cooperation for an ongoing data sharing arrangement. Second, when entering a data sharing arrangement there is limited oversight into the supplier’s processes. This lack of oversight tends to increase over of the various sources of data and gives rise to an increasing uncertainty about the data’s quality. Lastly, the statistical agency’s role will transition from directly managing data collection efforts to performing batch review of acquired data. Most concerns may be mitigated by carefully considering if the data should be collected and how it will be used. Other concerns may be addressed by helping the data supplier improve their processes.

4. Quality Assessment Tool

When developing this tool we considered the circumstances presented and made a list of questions for each dimension. We soon realized that the number of users, information products, suppliers, and data sources made it impractical to construct a unique checklist for every situation.

As a result, we propose a methodology that can be adapted to various circumstances. This methodology recognizes that in order to assess quality one must compare users expectations with actual observations. The basic structure for the tool therefore poses various questions that require the “Statistical Agency” (S) to document their quality expectations and the “Administrative Agency” (A) to document the actual characteristics of their data or processes. The Statistical agency would then document their assessment of the data’s fitness for the intended use.

As development of the tool progressed, the number of questions for each dimension accumulated. When we looked over all the questions there were some concerns about the response burden. To minimize the effort involved, we prioritized the usefulness of each question and segmented the questions into three phases for when the information would be needed. The first phase is for discovery. In this phase the statistical agency asks ten questions to gather the minimum amount of information to determine if an existing data source would be worth acquiring. Most of these questions are in the relevance and accessibility dimensions. The second phase is for initial acquisition of data. In this phase the statistical agency is sure the data can be of use and that it would be worthwhile to establish a data sharing agreement. At this point responses for the remaining quality dimensions are collected. This phase asks 24 questions and requires the most effort but ideally should be a one-time occurrence. The third phase is for repeat acquisition. In this phase the statistical agency continues to get periodic updates from the supplier. To minimize the burden in completing responses for this phase, we ask eight questions primarily geared at identifying changes from prior acquisitions. Most of these questions focus on coherence and accuracy.

Overall, the tool has 36 questions with six questions reused between initial and repeat acquisition. The relevance dimension has seven questions occurring in the discovery phase and primarily focuses on documenting concepts, coverage, entities, and timeliness. The accessibility dimension has four questions split between the discovery and initial acquisition phases. These questions document the legal requirements, costs, confidentiality, and medium by which the data would be transferred. The interpretability dimension has five questions mostly in the initial acquisition phase. These questions document the data sources, the modes of data collection, the questions used to originally acquire the data and any associated instructions, and the architecture used to store the data including a data dictionary. The coherence dimension has five questions in the initial acquisition phase with most being reused in the repeat acquisition phase. These questions document the taxonomies used and comparability of any aspect of the data program over time and geographic location. The accuracy dimension has seven questions all in the initial acquisition phase with two being reused in the repeat acquisition phase. These questions attempt to illuminate a true response and document a variety of possible errors that could be associated with the data. The institutional environment dimension has eight questions with all but one in the initial acquisition phase and the last in the repeat acquisition phase. We intended these questions to document the integrity of the data supplier and primarily focus on what quality systems they have in place and the effectiveness of those systems.

Maintaining documentation from this tool helps ensure quality in a number of ways. First, the statistical agency can help ensure that some critical requirements are not overlooked. Second, by informing the data supplier about the agency’s information needs and how the data will be used the statistical agency can begin building collaborative relationships with data suppliers. Such collaboration can help the data supplier discover other relevant data and may help facilitate data access by allaying any legal and privacy concerns on how the data will be used. Third, documenting a review of acquired data against user needs can help ensure that data is used in an appropriate manner. Fourth, documentation can be useful for the statistical agency to inform staff about the background and current status of data acquired from a given supplier. The data supplier can also use any analysis from the statistical agency as a feedback loop to improve the quality of their processes. Last, the documentation helps provide traceability to what data was collected and where it was used. This traceability can be of considerable importance in the event that one discovers an error at some point in the future.

5. Implementation

When using this tool, it is possible that a data supplier could provide multiple data files or even tables from a database. This tool should be flexible enough to document any situation. It is however important to make sure that each response is clearly associated with specific data files. A conscious decision should be made to either complete the tool for an entire collection of data files with clarifications made within a given responses to highlight any unique characteristics of a specific file or to complete a separate tool for each file.

The assessment tool has been pilot tested with a number of State, Federal, International, and Third Party organizations. We solicited their feedback to find out if they felt the tool would be useful to assess quality and get their ideas on making the tool more practical to use. Based on the results, we rarely got an immediate response to our request to complete the tool. Given the breadth of questions asked, the data supplier typically needs to contact a number of people in order to get responses for all of the questions. In some instances, legal review delayed responses from data suppliers. However, we do predict that a data supplier might take about six to ten hours to complete the discovery and initial acquisition phases—if a data dictionary is already available.

Overall, the tool is viewed as providing useful information. Data suppliers did voice some concerns about possible misinterpretation of data quality assessments. Data suppliers are concerned that their data is labeled as bad quality when in reality it is just not suitable for a statistical agency’s intended use. Statistical agencies expressed a desire to make it easier to arrive at a quality recommendation. One suggestion included prioritizing the questions and another expressed an interest in some method to calculate a quality score.

6. Recommendations

This process will work best in an infrastructure that promotes transparency. Before you can acquire data, you have to know that it exists. For the United States, the Office of Management and Budget has made efforts in this area by issuing an executive memorandum for an “Open Data Policy – Managing Information as an Asset.”[[[10]](#footnote-10)] This effort establishes some minimum metadata requirements that will facilitate a statistical agency’s ability to discover if an existing records would be relevant to replacing a fielded question or could be used to enhance an existing information product.

Collaborate with data suppliers. Since you require with the same data, you should have shared interests in the information product. Cooperation with a data supplier can help ensure a mutually beneficial and ongoing relationship. Completing this assessment tool can be instrumental in helping the statistical agency provide a data supplier feedback on the quality of their data as well as offer assistance in improving their processes.

Maintain your documentation. The records you keep will be useful in keeping staff informed about the status of any acquired data. Furthermore, if an error is found at some future point the documentation you maintain can give you some traceability to finding affected data sources and information products.

| Expectation | Acquisition | | | | Assessment | Data Quality Assessment Tool – May 2014  Compendium of Questions  **Questions (By Phase, Dimension, and Responding Agency)**  (S – Statistical Agency, A – Administrative Agency) |
| --- | --- | --- | --- | --- | --- | --- |
| Discovery | Initial | Repeat | |
| **Relevance.** Relevance refers to how well the administrative data meets the needs of the user in regards to data level (person, family, household, establishment, company, etc.), broad data definitions and concepts, population coverage, time period and timeliness. | | | | | | |
| S | A |  | |  | S | **S:**  What benefit and products should be derived from using the desired data?  **A:**  What is the purpose of the administrative program? |
| S | A |  | |  | S | **S/A:** What scope of records are needed / exist for the administrative data (characteristics of program participants, geographic areas, etc.)? |
| S | A |  | |  | S | **S/A:**  Does the dataset contain records for those who applied for but were are denied eligibility to the program or benefits? |
| S | A |  | |  | S | **S/A:** What record level entity is needed / exists for the administrative data (person, business, etc.)? |
| S | A |  | |  | S | **S:**  How frequently will statistical reports be published? **A:** How often are the data collected at the Administrative Agency? (Continuously, Monthly, Quarterly, Annually, etc.)? |
| S | A |  | |  | S | **S/A:** What reference time period is needed / available for the administrative data? |
| S | A |  | |  | S | **S/A:** How soon after a reference period ends is data needed / can data be provided? |
| **Accessibility.** Accessibility refers to the ease with which the data can be obtained from the administrative agency. This includes the suitability of the form or medium for transferring the data, confidentiality constraints, and cost. | | | | | | |
| S | A |  |  | | S | **S/A:**  Describe any legal, regulatory, or policy requirements permitting / restricting access to the data. |
| S | A |  |  | | S | **S/A:** What is the typical reimbursement given / received for sharing the data? (Are reports, a data product, or a monetary fee given in exchange?) |
| S |  | A |  | | S | **S:**  Describe variables of particular interest to the statistical program. **A:**  Describe variables that were collected but cannot be included (for confidentiality or other reasons) in a data file extract or that may only be included in a modified form (i.e. Replacing names with identification numbers). |
| S |  | A |  | | S | **S/A:** What method is preferred to access the data? (i.e., Remote access to a database, Transferring a data file to a cloud server, Mailing a data file on a DVD.) |
| **Interpretability.** Interpretability refers to the clarity of information to ensure that the administrative data are utilized in an appropriate way. This includes evaluation of data collection forms, data collection instructions, and a data dictionary. | | | | | | |
|  | A |  |  | | S | **A:**  Provide a copy of any forms and instructions used to collect data from an applicant. |
|  |  | A |  | | S | **A:**  Who is source data collected from? (i.e., Self-reported, Proxy for respondent, Linked records) |
|  |  | A | A | | S | **A:**  Describe each variable on the administrative data file and note their valid values. This also includes metadata for denied applicants, recoding, and revision. (See data dictionary template) |
|  |  | A |  | | S | **A:**  Describe the architecture for storing the data and relationships between key variables. For example, are data stored in a spreadsheet with one row for each person/entity, a relational database, or some other format? |
|  |  | A |  | | S | **A:**  How is source data collected? (i.e., paper questionnaire, computer assisted person interview, computer assisted telephone interview, web data collection form) |
| **Coherence.** Coherence refers to the degree to which the administrative data are comparable with other data sources and consistent over time and across geographical areas. This includes evaluation of data concepts, classifications, questionnaire wording, data collection methodologies, the reference time period, and the target population. | | | | | | |
| S |  | A |  | | S | **S/A:** Describe any taxonomy used to categorize or classify the data. For example: Do questions asked about industry classification use the current North American Industry Classification System (NAICS)? |
|  |  | A | A | | S | **A:** Describe any changes to or differences in the questions asked and any changes to instructions for completing questions. (i.e., posing new questions, using different wording for the same question, revising questions or instructions, introducing new languages, or deleting questions) |
|  |  | A | A | | S | **A:** Describe any changes made to how the data are processed. (i.e., changes to mode of data collection, changes to edits, changes to classification codes, or changes to the query system used to retrieve the data) |
|  |  | A | A | | S | **A:** Describe any changes that influenced who participated in the program. (i.e., legislative changes, changes to eligibility requirements, changes to geographical boundaries**,** or occurrence of natural disasters affecting program participation.) |
|  |  |  | A | | S | **A:** Describe any new records or revisions to existing records that may occur after data acquisition and when those changes are likely to occur. |
| **Accuracy.** Accuracy refers to the closeness of the administrative record data values to their (unknown) true values. This includes information on any known sources of errors in the administrative data such as missing records, missing values of individual data items, misinterpretation of questions, and keying, coding, and duplication errors. | | | | | | |
| S |  | A |  | | S | **S/A:** What percentage of those who are eligible/mandated to apply/report are not included on the data file. What is known about their characteristics? |
| S |  | A |  | | S | **S/A:** What are the known sources of errors in the administrative data (e.g. non-response, keying, coding errors) and the magnitude of their effect? |
| S |  | S | S | | S | **S:** What is the range of values and frequencies of responses (including proportion of missing values) for each field in the administrative data file? (Include attachments for documentation.) |
|  |  | A |  | | S | **A:** Describe quality control checks or analyses and the typical results for your production processes. |
|  |  | A |  | | S | **A:** What questions are most often misinterpreted? |
|  |  | A |  | | S | **A:** Which items are subject to revision either by editing or updating data values? What fields are revised the most? |
|  |  | S | S | | S | **S:** What are the record counts and proportion of duplicate records in the administrative data file? |
| **Institutional Environment.** Institutional Environment refers to the credibility of the administrative agency for producing high quality and reliable administrative data. This includes an evaluation of the agency’s quality standards and processes for assuring adherence to standards. | | | | | | |
| S |  | A |  | | S | **S:** Describe the statistical agency’s expectations for the administrative agency’s quality standards. **A:** Describe the administrative agency’s quality standards. |
| S |  | A |  | | S | **S/A:** Provide contact information for managerial and technical staff that may help address any questions or concerns about use of the data. |
|  |  | A |  | | S | **A:** Describe the administrative agency’s data processing from data collection through editing, review, correction, dissemination, and retention. |
|  |  | A |  | | S | **A:** Describe the administrative agency’s process to control changes to their data processing. |
|  |  | A |  | | S | **A:** Describe any audit records that document changes made to data values from what was originally reported,. These records typically explain such things as who changed the value, what the original value was before the change, when the value was changed, why the value was changed, and how the value was changed (e.g., programmatic edit or manual correction). |
|  |  | A |  | | S | **A:** Describe the findings and corrective actions resulting from studies, evaluations, or audits to assess your organizations adherence to its quality standards. |
|  |  |  | A | | S | **A:** Describe corrective actions taken to improve the quality of your processes and data. |
|  |  | A |  | | S | **S/A:** Document the Name, Phone #, and E-mail address for managerial and technical contacts. |

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Examples** |
| File Name | Provide the filename of the table | Applicants.txt  Providers.csv |
| File Date | Date the file was saved. | 3/4/12 |
| Description | Enter a description of the file contents indicating the file extract period for the information | 2004 program beneficiaries |
| File Format | Describe how the data is stored on the file. | Comma separated, Quote delimited values |

**For each column in the table, provide the following information:**

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Examples** |
| Field Name | Name of each field/column as it appears in the table | FName, LName, Address, Income, |
| Description | Description of the field/column. | "Applicants First Name" |
| Type | Data type of the fields/columns in the table and field/column size | Text, Character, Number, Integer, Double, Yes/No,  Date/Time, Currency, Hyperlink |
| Req. | Is a response required? | Y, N |
| Len | The length of the field. | 8.2, 20, 256 |
| Format | The format that the data is stored in the table. | ###-##-####,  (###) ###-####  ###.## |
| Units | Physical units for the value of the field/column. Leave blank if not relevant. | Hours, Dollars, Acres, Thousands |
| Valid Values | The values that would be accepted as a valid response. | Null  "1"  "2" |
| Definitions | Define the meaning for the codes that are used. Identify which value is the default value | Null – No response (Default)  1 - Active  2 - Denied |
| Notes | Note any other items of special consideration.   * Is this field a primary or secondary key? * Is this field indexed? * Are the valid values and definitions based on a recognized standard like ZIP codes? If so, note the standard and version used. * Note other considerations. | Primary Key  Indexed  USPS ZIP Code Standard 1/1/05 |

1. [] Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau. [↑](#footnote-ref-1)
2. [] Office of Management and Budget (09/2006) Standards and Guidelines for Statistical Surveys. <http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf> [↑](#footnote-ref-2)
3. [] Office of Management and Budget (02/2014) Memorandum M-14-06, Guidance for Providing and Using Administrative Data for Statistical Purposes. <http://www.whitehouse.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf> [↑](#footnote-ref-3)
4. [] U.S. Census Bureau (June 2012) Statistical Quality Standards. <http://www.census.gov/quality/standards/index.html> [↑](#footnote-ref-4)
5. [] Statistics Canada (10 April 2014). Statistics Canada Quality Guidelines

   <http://www.statcan.gc.ca/pub/12-539-x/2009001/introduction-eng.htm> [↑](#footnote-ref-5)
6. [] Australian Bureau of Statistics (05 May 2009) ABS Data Quality Framework <http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1520.0Main%20Features1May%202009?opendocument&tabname=Summary&prodno=1520.0&issue=May%202009&num=&view>= [↑](#footnote-ref-6)
7. [] Office for National Statistics (Sep 2013) Guidelines for Measuring Statistical Quality. <http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/index.html> [↑](#footnote-ref-7)
8. [] European Statistical System. (2012) Quality Assurance Framework of the European Statistical System <http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/QAF_2012/EN/QAF_2012-EN.PDF> [↑](#footnote-ref-8)
9. [] Eurostat. (03 Oct 2003) Quality Assessment of Administrative Data for Statistical Purposes <http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/QUALITY%20ASSESSMENT%20OF%20ADMINISTRATIVE%20DATA%20FOR%20STATISTICA.pdf> [↑](#footnote-ref-9)
10. [] Office of Management and Budget. (09 May 2013) Memorandum M-13-13 Open Data Policy – Managing Information as an Asset. <http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf> [↑](#footnote-ref-10)