Data Quality Assessment Tool for Administrative Data

William Iwig
National Agricultural Statistics Services

Michael Berning
U.S. Census Bureau

Paul Marck
U.S. Census Bureau

Mark Prell
Economic Research Service

February 2013

Contents

Data Quality Assessment Tool for Administrative Data

William Iwig, Michael Berning, Paul Marck, Mark Prell[1]

Introduction

Government agencies maintain data for the administration of government programs. Unlike survey data, which are collected for statistical purposes, administrative data are collected as part of a program agency's routine operations. Through a Memorandum of Understanding (MOU) a program agency may share its administrative records, or a data file extract of those records, with a statistical agency (under provisions that protect data confidentiality). The extent and effectiveness of data-sharing are impeded by various barriers. One of these barriers is limited information on how to assess the quality of the data.

Broadly defined, data quality means "fitness for use." Different users of the same data can have different assessments of its quality. Administrative data were gathered for a particular purpose—running a program—and can have qualities that are well-suited for that purpose. When the data are adapted to a new purpose, issues of data quality become especially salient.

Assessment of data quality can benefit both program and statistical agencies. Statistical agencies that seek more routine use of administrative records in their statistical programs benefit from understanding the quality of the administrative data they receive. Upon receipt, a statistical agency produces various reports, analyses and data products that may be informative for program managers, policy officials and the program's various stakeholders. The benefits to a program of these statistical analyses can be substantial. In addition, the program agency may face the need to increase understanding of its administrative records system among its own staff. As senior staff transition to different jobs or retire, ongoing staff responsible for managing the agency's data system can benefit from fuller understanding and more complete documentation of key aspects of the system including the quality of the data.

In July 2011, the Interagency Council on Statistical Policy asked the Statistical Uses of Administrative Records Subcommittee (of the Federal Committee on Statistical Methodology (FCSM)) to examine the

[1] This paper represents part of the ongoing work of an interagency subcommittee under the auspices of the Federal Committee on Statistical Methodology (FCSM). The views expressed represent the individual authors and not necessarily the full FCSM or the agencies at which the authors are employed, which are: Economic Research Service (Mark Prell), National Agricultural Statistics Service (William Iwig) and the U.S. Census Bureau (Michael Berning, Paul Marck).

quality of administrative data gathered by government agencies. The Subcommittee's Data Quality Working Group had already begun foundational cross-agency work that developed into the current *Data Quality Assessment Tool for Administrative Data* presented here (hereafter referred to as the Tool). This Tool is developed to support a conversation between a *user* of an agency's administrative data —either a user who may be initially unfamiliar with the structure, content and meaning of the records or a user who repetitively acquires the data but may not be aware of recent changes to the system—and a knowledgeable *supplier* or provider of administrative data. The Tool provides questions that are pertinent to helping a user assess the fitness for their intended use. These broad questions can prompt more particular, data-specific questions in the discussion between the data receiver and the data provider.

The Tool is designed to help promote data quality in two ways. First, by helping a user better understand the data's attributes, the Tool enables the data to be used more appropriately and meaningfully for their new applications—ones that differ from their original administrative uses. Second, the Tool can help promote data quality over time, both for administrative purposes and statistical purposes. While the Tool does not describe all the steps needed for improving quality, the Tool can help a data provider evaluate current quality conditions. Which aspects of data quality are already strong? In what ways? Which other aspects of data quality could benefit from strengthening? Assessing and understanding *current* quality is the critical first step by which a data provider would begin a long-term process by which to *improve* data quality over time.

A key feature of data quality assessment that is built into this Tool is the recognition that data quality is not homogenous but instead has several *dimensions* (or "characteristics" or "features"). This multi-dimensional structure is a common feature of the data quality frameworks for other national statistical offices. To draw from their expertise and build on their previous work, the Data Quality Working Group examined the frameworks of the Australia Bureau of Statistics[2], the United Kingdom's Office of National Statistics[3], Statistics Canada[4], and Eurostat[5]. While these frameworks each focus on the quality of statistical products and output, they do differ to some degree in scope, terminology, and application. Some frameworks were designed to be applied for the various steps in the statistical production process, including the use of administrative data sources. Based on this review, the Working Group developed a

[2] ABS Data Quality Framework, May 2009 (http://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0)
[3] ONS Guidelines for Measuring Statistical Quality (http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/index.html)
[4] Statistics Canada Quality Guidelines (http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.htm)
[5] European Statistics Code of Practice (http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF)

tool containing six quality dimensions: Relevance, Accessibility, Coherence, Interpretability, Accuracy, and Institutional Environment. The Tool is specifically designed to assess the quality characteristics of an administrative data source. It provides a definition of each of its six dimensions and a set of questions that help the data receiver and the data provider evaluate the quality of an administrative data source relative to each dimension. Some of the dimensions overlap but each question is only included under one dimension. Timeliness is a common separate dimension in many of the prior frameworks but it is covered here under the Relevance dimension of the Tool. Coherence and Comparability are frequently included as separate dimensions in other frameworks but are both included under the Coherence dimension in the tool, covering consistency over time periods and comparability over domains.

Using the Tool does not result in a single overall numerical measure (or metric or index) for data quality. Instead, the Tool provides questions for which some answers are quantitative and others qualitative. The usefulness of the Tool lies in providing a well-developed set of questions that prompt the user to consider certain key attributes of data quality; the Tool does not result in a judgment or recommendation apart from what the user develops. It is the user's own interpretation of the answers—and the user's prioritization of which ones are especially germane for the data application at hand—that constitutes the user's own assessment of data quality. .

The Tool contains 43 questions organized by three "phases" of the data sharing process. Then, within each phase, the questions are organized by the dimensions of data quality that are relevant to that phase. Consequently, only a subset of the questions must be answered at any one time. The 12 questions in the initial *Discovery* phase cover information needed to determine the feasibility and desirability of a possible data sharing project and ultimately, to approve the development of an MOU. The data quality dimensions of Relevance, Accessibility, and Interpretability are central to the Discovery phase. The *Initial Acquisition* phase begins with the approval for developing an MOU, includes finalizing and signing of the MOU, and ends with the first-time receipt of the data. During this phase, more detailed information will be needed by statistical staff on data quality issues, such as recoding methods, known sources of errors, and proportion of missing values for each data field. The data quality dimensions of Accessibility and Interpretability are important in this phase, as they were in the Discovery phase, each with new questions that arise with Initial Acquisition. This phase also has questions for the dimensions of Coherence, Accuracy, and Institutional Environment. Altogether this phase has 29 questions. Questions in the *Repeated Acquisition* phase cover periodic receipt of data. These questions would not be answered unless and until a second installment of the data is planned. The Repeated Acquisition phase has 11 questions, two of which are new while nine are repeated from the Discovery or Initial Acquisition phase.

Questions listed in the Tool under one phase may, in an actual instance, be helpful in a different phase. The ordering or sequencing of questions in the Tool need not be binding. **Appendix A—Compendium of Questions Across All Three Phases** provides a table containing the Tool's full set of questions for easy reference. While the Tool contains 43 questions, only some of the questions are answered in any one of the three phases. Those questions that are not helpful for a user are ignorable, while other questions may prompt a user to think of issues or questions in addition to those provided by the Tool.

Based on pilot test results, the time required for the staff of the data provider to complete the Tool (for the early phases) is predicted to be about six to ten hours—if a data dictionary is already available. If a data provider does not have a data dictionary, then a much larger but unknown amount of time would be involved with preparing one. The Tool's **Appendix B—Data Dictionary Template** provides an example of a core set of information that is typically contained in a data dictionary.

As suggested previously, good communication between the program agency and the statistical agency regarding the quality characteristics of the administrative data file is very important. Depending on the formality of the data sharing arrangement, communication protocols and contact persons may be specified in the data sharing agreement. Both agencies will likely continue to discover new findings about the quality of the data during the data sharing arrangement which need to be communicated to the other agency.

To facilitate completion of the Tool, a hypothetical example is provided with *"Example answers"* involving files from a generic State agency administering the Supplemental Nutrition Assistance Program (SNAP, formerly the Food Stamp Program). SNAP is a federal-state partnership administered at the federal level by the Food and Nutrition Service of the U.S. Department of Agriculture, and at the State level by separate State SNAP agencies.

In SNAP, the group of persons who usually purchase and prepare foods together is known officially as the "SNAP unit." The example below substitutes the more common term "family."

When using this Tool, it is important to recognize that the number and content of the *administrative data files* maintained in a program agency's data system may differ from the number and content of the *data file extracts* that are derived or taken from the agency's administrative data files and sent to the data-receiving statistical agency to conduct a research project. The Tool is flexible enough to be completed:

- once, by discussing each separate file as needed in the answer; or
- separately for each separate file.

What is key is that a conscious decision be made to adopt one of those two approaches, that the answers be written clearly for that approach, and that (if necessary) clear distinctions be provided between the administrative data files and the data file extracts. The hypothetical example we provide for SNAP happens to involve *two* administrative files. Each file is discussed, as needed, in the answers in the Discovery phase in anticipation of sending *two* data file extracts, one from each original administrative file, in the Initial Acquisition phase. The example also includes detailed data dictionaries for the Discovery and Repeated Acquisition phases.

I. Discovery Phase

Questions in the Discovery phase cover information needed to approve the development of an MOU. The data quality dimensions of Relevance, Accessibility, and Interpretability are central to the Discovery phase. **(12 questions)**

RELEVANCE. Relevance refers to how well the administrative data file 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**.**

1. What are the general contents of the administrative data file(s) maintained by the program?

Example answer. In our SNAP data system, we organize data into two administrative data files. Administrative File 1 is a "Program Beneficiaries" file that contains address information for the family. Administrative File 2 is a "Benefits" file that contains information on the amount of SNAP benefits for the family and personal information about each of the family's members (e.g., birth date, gender, etc.). If we were to complete an MOU with you, it would be convenient for us to prepare two data files extracts, one for each of our original administrative files.

2. What entity is represented at the record level of the administrative data file (person, household, family, business, etc.)?

Example answer. Administrative File 1 is organized by family (the Benefit Case Number is a family's unique identifier). Administrative File 2 is organized by individuals receiving benefits (the Case Number, Benefit Month, and Client ID serve as a unique identifier). The data file extracts would be organized the same way.

3. Describe the scope of the records included on the administrative data file (program participants, geographic areas, etc.).

Example answer. The administrative files contain data for those who officially begin the application process, including (a) applicants who are determined to be eligible for benefits and become SNAP participants for some period, (b) applicants who are ineligible for SNAP benefits and are not participants, and (c) applicants who do not complete the application process. The application process is officially begun with the submission of a SNAP application form that contains the names of the persons in the family, a signature, and an SSN. The geographic scope covers all applications filed in any program office in the State. If you want a data file extract for a particular sub-State area, or you want data fields (zip codes) that indicate such areas, the MOU should specify.

4. What are the earliest and latest dates for which administrative data are available (including in archive)?

Example answer. Our system contains active data for the calendar year-to-date and for the 5 previous calendar years. As each year ends, the earliest year of active data is moved to archive where they are retrievable, although with difficulty. Archives extend to 1998. Data earlier than 1998 have been purged and are not retrievable.

5. How often are the data collected at the program office? (e.g., Daily for flow data, or a particular date such as April 1)?

Example answer. Daily, at least for days when program offices are open.

6. What is the "reference time period" by which the data are organized in the administrative data file?

Example answer. Although data are collected daily (question 5), in our State's database there is not a separate file for each business day. Instead, we organize the data by month in our administrative data system. If an MOU is written, it would be good if it specified whether you want the data in your data file extract to be organized by month, or if it should be aggregated to the quarter or the year before we send it to you.

7. How soon after a reference time period ends can a data file extract be prepared and provided?

Example answer. After the end of a month, about 3-4 weeks would be needed to prepare a data file extract that includes that months' data.

ACCESSIBILITY. Accessibility refers to the ease with which the data file extract can be obtained from the administrative agency. This includes the suitability of the form or medium for transferring the data, confidentiality constraints, and cost.

8. Describe any legal, regulatory or administrative restrictions on access to the data file extract.

Example answer. There are two restrictions on any data file extract. First, if an MOU is written by which the data can be shared, the MOU will describe the project(s) for which the data can be used with the restriction that the data cannot be used for any other project (unless the MOU is amended). Second, the data are to be kept confidential. The MOU, or a supplementary Data Security Agreement, would need to specify the technical and procedural safeguards by which the agency that receives the data will limit access to the data and maintain confidentiality. In particular, the MOU would need to describe how Personally Identifiable Information (PII) will be protected and how access to the PII will be permitted only to authorized persons.

9. Describe any records or fields that were collected but cannot be included (for confidentiality or other reasons) in the data file extract provided to the user.

Example answer. There are no restrictions on the availability of data fields, provided they are specified in the MOU.

10. What is the typical reimbursement for sharing your data? Are reports, a data product, or a monetary fee given in return?

Example answer. In exchange for providing data access to the statistical agency, the State SNAP agency requests that the statistical agency provide summary reports showing the results of analysis that includes the State SNAP file and tabulations quantifying the quality of the State administrative records data.

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.

13. Describe each variable on the administrative data file and note their valid values. (See template for data dictionary).

Example answer. See attached data dictionary.

14. If a complete data dictionary is not available, describe the primary identification and response fields on the administrative data file and note their valid values.

Example answer. See attached data dictionary.

Example Data Dictionary for Discovery Phase

File Name:	Program Beneficiaries.por	Description:	2004 Beneficiary Contact Information
File Date:	February 14, 2005	File Format:	SAS portable dataset

Field Name	Description	Req.	Type	Len	Format	Units	Valid Values	Definitions	Notes:
case	Case ID	Y	Num	6			1 - 999999	Sequential number assigned to each case	Primary Key
yrmo	First benefit month case appeared at address	Y	Num	4	yymm		04 01 - 12	yy – two digit year mm – two digit month	
resaddr	Residential address	Y	Char	1			Null Y N	No response **(Default)** Yes this is a residential address No this is not a residential address	
HHName	Name for household	Y	Char	40			Alphabetic		
addrln1	Mailing address – line 1	Y	Char	22			Alpha-numeric		
addrln2	Mailing address – line 2	N	Char	22			Null Alpha-numeric	No response **(Default)**	
city	City	N	Char	16			Alphabetic		
St	State	N	Char	2			Alphabetic	See Notes	Definitions are contained the USPS standard for ZIP codes dated Jan 1, 2005
Zip	5 digit zip code	Y	Num	4			See Notes	See Notes	Valid values and definitions are contained in USPS standard for ZIP codes dated Jan 1, 2005
fips_county	FIPS County code	N	Num	3			See Notes	See Notes	Valid values and definitions are contained in FIPS standard dated Oct 24, 1998.

Example Data Dictionary for Discovery Phase

File Name:	Benefits_2004.por	Description:	Benefits data for 2004.
File Date:	February 14, 2005	File Format:	SAS portable dataset

Field Name	Description	Req.	Type	Len	Format	Units	Valid Values	Definitions	Notes:
case	Case ID	Y	Num	6			0 - 999999	Sequential number assigned to each case	Primary Key
yrmo	Benefit month	Y	Num	6	yyyymm		2004 01-12	yyyy – 4 digit year mm – 2 digit month	Primary Key
CID	Client ID	Y	Char	6			0 - 999999	Sequential number assigned to each client	Primary Key
client	Name of client	Y	Char	40			Alphabetic	Full name of client	
ssn	Social Security Number	N	Num	9	###-##-####		Null Numeric	**(Default)**	
birthdt	Birthdate	Y	Num	6	mmddyy		01-12 01-31 00-99	mm – 2 digit month dd – 2 digit day yy – 2 digit year	
race	Ethnic Group	N	Char	1			1 2 3 4 5 6	White Black Hispanic American Indian or Alaskan Native Asian or Pacific Islander Omitted **(Default)**	
sex	Client gender	N	Char	1			M F U	Male Female Unknown **(Default)**	
sig1	Code which identifies the client's status in group	Y	Char	1			A B C D E F G H J K M	Head of Household member Student Striker Military member Eligible legalized alien Treatment center residents Head of Household - nonmember Refugee Disqualified as ineligible alien Disqualified failed to meet SSN requirement	

Field Name	Description	Req.	Type	Len	Format	Units	Valid Values	Definitions	Notes:
							R S T U X W Z	Migrant, out of work stream DHS employee Alien with acceptable alien status Disqualified for intentional program violation Seasonal farm worker Non primary wage disqualified for noncompliance Migrant, in work stream Alien granted temporary resident status	
educat	Client Education Level	Y	Char	1			1 2 3 4 5 6 7 8 9 A B C D E F N	First Grade Second Grade Third Grade Fourth Grade Fifth Grade Sixth Grade Seventh Grade Eighth Grade Ninth Grade Tenth Grade Eleventh Grade High School Graduate/ GED Currently Attends JR, High, GED Attending or completed some college Graduate of a 4 yr college No Formal Education	
casestat	Case status code	Y	Num	3			1 2 3 4	Active Denied Suspended Hold **(Default)**	
coupon	Benefit amount	Y	Num	3	$###	Dollar	0 - 999	$0 (Min) **(Default)** $999 (Max)	
serlvl	The client's time limited benefits tier level	Y	Char	1			1 2	Likely to obtain above min. wage within 6 mo. Can obtain above min. wage employment within 1 yr with case	

Field Name	Description	Req.	Type	Len	Format	Units	Valid Values	Definitions	Notes:
								management	
							3	Client has severe impediment to employment	
							4	Client refused to cooperate	
							5	Client has 18 months of work experience or completion of HS, GED, Post Sec., Tech or Voc. School	
							6	Client has 6-17 months recent work experience or completion of 11th grade	
							7	Client has < 6 months of work experience and completion of less than 3 years of	
							8	Plucker assigned default for SIG 7,8 clients	

II. Initial Acquisition Phase

The Initial Acquisition phase begins with the approval for developing the MOU, includes the finalizing and signing of the MOU, and ends with the first-time receipt of the data. The data quality dimensions of Accessibility and Interpretability are important in this phase, as they were in the Discovery phase, each with new questions that arise with Initial Acquisition. This phase also has questions for the dimensions of Coherence, Accuracy, and Institutional Environment. **(29 questions)**

ACCESSIBILITY. Accessibility refers to the ease with which the data file extract can be obtained from the administrative agency. This includes the suitability of the form or medium for transferring the data, confidentiality constraints, and cost.

11. Describe any disclosure control methods or other modifications of the administrative data file used to protect the confidentiality of the data (i.e., Topcoding, Replacing names with identification numbers), or will non-modified data from the program agency's database(s) be provided and accessible to the statistical agency?

Example answer. No disclosure control methods are used to suppress or alter the data when the data are drawn from the State's administrative data files and placed into data file extracts. The data file extracts are protected via an encryption key that will be provided to the statistical agency under separate correspondence.

12. How will the data file extract be transferred to the user?

Example answer. Because data file extracts are encrypted, they can be on DVD shipped via traceable mail.

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.

15. If a complete data dictionary is not available, describe any remaining fields (not already documented in the Discovery phase).

Example answer. Data Dictionary provided

16. Provide a copy of any forms and instructions to the applicant used to collect data.

Example answer. See our State's website at www.SNAPInstructions.GenericState.gov for a copy of the form used to collect data and the instructions.

17. Describe the methodology used to recode original data to create a value for a new variable, such as assigning a reported age to an age range.

Example answer. No recodes—data are keyed either from paper forms or as captured during interview.

18. Describe the data architecture and relationship 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?

Example answer. Data are extracted from a state database and written to DVD as two flat files. One file uses case id as a primary key and contains information related to the case such as the address for the family. The other file uses a composite case id-benefit month-client id as a primary key and contains data related to the individual client, benefit amounts, and payment dates.

19. Does the dataset contain records for those who are denied eligibility to the program or benefits and how can those records be identified?

Example answer. Yes, the SNAP database contains four types of cases, as indicated by the variable "casestat," which is the case status code. There are four codes: active, denied, suspended and on hold cases. If the statistical agency wants data only for a subset of cases in the State's database (e.g., only the "active" cases) then the MOU should specify which subset is wanted; otherwise it is likely to be presumed that the statistical agency wants data for all four types of cases.

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, reference period, and the target population.

20. Please describe any classification systems used for categorizing or classifying the data. For example: Do questions asked about race use the current race and ethnicity categories defined by the Office of Management and Budget (OMB)? Do questions asked about industry classifications use the current North American Industry Classification System (NAICS)?

Example answer. We use a classification system for race and ethnicity in a single database field. We combine 4 standard racial categories (White, Black, American Indian or Alaskan Native, Asian or Pacific Islander) with an ethnic category (Hispanic) into a set of 5 either/or alternatives. In this system, a person could be classified as, for example, either White or Hispanic but not both.

21. Were there changes made during the file extract period (e.g., January 2006 through December 2008) or differences across the geographical areas covered by the file in the questions asked that would cause a break in consistency, such as different questions, revised questions, questions in different languages, or deleted questions?

Example answer. Based on our conversations of what the file extract period is likely to be for your statistical study (October 2003- September 2007), there were no changes in this period for the examples you list.

22. Were changes made during the file extract period to how the data are processed, such as changes to mode of data collection, changes to instructions for completing the application form, changes to the edit, changes to classification codes, or changes to the query system used to retrieve the data?

Example answer. In 2005 the State introduced an option by which a SNAP applicant could directly enter (some) of the family's information on-line; a limited amount of paper-based information, including a signature, continued to be required of all persons submitting a SNAP application. The previous paper-based form (combined with computer-assisted personal interviews) continued to be available for those who did not choose to use the on-line option at all.

23. Were there changes during the file extract period to geographical boundaries?

Example answer. In June 2005, the geographic area for zip code 12345 was reclassified into zip codes 12344 and 12345. The records for the number of participants in zip code 12345 shows a sharp drop in June 2005.

24. Briefly, describe substantial changes during the file extract period or differences across geographical areas covered by the file that influenced who participated in the program, such as legislative changes, changes to eligibility requirements, expansions of the program, or natural disasters impacting program participation?

Example answer. Households where all members benefit from cash assistance provided through the Temporary Assistance for Needy Families (TANF) Program are categorically eligible for SNAP. State SNAP agencies have the option to expand categorical eligibility so that households receiving TANF services other than cash assistance are categorically eligible for SNAP. In January 2004, our State adopted the option of expanded categorical eligibility. As a result of adopting this option, the number of families in the State eligible for SNAP increased.

25. Were there any other changes during the file extract period or differences across the geographical areas covered by the file that would cause a break in consistency in regards to comparing data over time? If so, describe the change/difference and when/where it occurred.

Example answer. Yes. Described in question 24 above. The number of families in SNAP could be higher following January 2004 than it would have been.

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.

26. What investigations/analyses have been conducted that reveal data quality characteristics (such as Government Accountability Office reports, Office of Inspector General audits, internal agency reviews, etc.)?

Example answer. SNAP agencies use identification (such as driver's license or photo I.D.), wage stubs, and other documents to check on personal, earnings and residential information reported by the family applying for SNAP. Is the person really who the person claims to be? Does the employer's report of the person's income agree with the person's report? These questions and SNAP agency procedures reflect one type of data accuracy that can fall under your category of "analyses." Let us know if you need more detail on the documentation we require from SNAP applicants.

A separate type of data accuracy is the accuracy of State eligibility and benefit determinations which is monitored by the SNAP Quality Control (QC) System. See materials on FNS website about QC at http://www.fns.usda.gov/snap/qc/default.htm. Payment error rates are calculated at the national and State levels. In Fiscal 2004, the national payment error rate was 5.88 percent including both overpayments of 4.48 percent and underpayments of 1.41 percent.

27. What percentage of eligibles are not included on the data file or what percentage of those mandated are not compliant? What is known about their characteristics?

Example answer. An annual report issued by Food and Nutrition Service provides estimates at the State level for participation rates among eligibles. The national rate in Fiscal 2004 was 61.8 percent (http://www.fns.usda.gov/ORA/menu/Published/SNAP/FILES/Participation/Trends2002-09.pdf.

28. What is the proportion of duplicate records on the data file extract?

Example answer. None. The file extracted for benefit data contains a composite primary key of Case ID-Benefit Month-Client ID that ensures one unique occurrence for each combination of these fields. There are regular quality control processes that prevent a single client from being assigned to multiple client ID's and a single Client ID's from being assigned to multiple Case ID's for a given benefit month .

29. What is the proportion of missing values for each field? (Feel free to provide an attachment.)

Example answer. Ethnic group is the primary field with missing values. Item nonresponse for the ethnic group field in fiscal year 2004 is 72%. The reason for this high percentage of missing values is that SNAP applicants can voluntarily respond whether they want to answer race/ethnicity questions (that is, an answer is not required in order to receive SNAP benefits)

30. Describe other metadata provided with the administrative data file such as record counts, range of values, and frequencies of responses.

Example answer. Metadata will include record counts and a data dictionary. The frequency of missing values for each field in the fiscal 2004 data is attached.

31. What are the known sources of errors in the administrative data (e.g. non-response, keying, coding errors)?

Example answer. Item nonresponse for the race/ethnicity variable in fiscal year 2004 is 72%. Keying errors are known to be less than 1%.

32. What questions are most often misinterpreted?

Example answer. In cases of joint custody there is often confusion about household composition and which Case ID a client should be assigned.

33. Which items are subject to revision either by editing or updating data values? What items are revised the most?

Example answer. Case status can change. Also, addresses are updated as people move and old addresses are overwritten.

34. If a value is revised from what was originally reported, describe any flags in the dataset that would indicate that the value was changed as well as explain why the value was changed.

Example answer. No flags are set. Change in case status can only be noted by comparing the case number across file months.

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.

35. Describe the purpose of the administrative program.

Example answer. The purposes of SNAP are to strengthen the agricultural economy; to help to achieve a fuller and more effective use of food abundances; and to provide for improved levels of nutrition among low-income households through a cooperative Federal-State program of food assistance to be operated through normal channels of trade.

36. Describe your processes for data collection, editing, review, correction, dissemination, and retention.

Example answer. Data collection is done via a client interview with periodic quality control from first-line supervisors. Clients sit face to face with an intake worker while the intake worker keys in the responses. Editing a case file after the fact can only be done with management concurrence. A 10-percent sample of cases is reviewed by management on a daily basis to determine the completeness of data. Data is maintained in the system for five years after which non-active files are archived.

37. Who is source data collected from (self-reported, interview a third party)?

Example answer. The client or someone representing the client.

38. How is source data collected (paper questionnaire, computer assisted person interview, computer assisted telephone interview, web data collection form)?

Example answer. A combination of paper questionnaire and computer-assisted person interview

39. Describe the checks the administrative agency performs to ensure the quality of the data and the typical results for your production processes.

Example answer. See our State's website at www.SNAPInstructions.GenericState.gov/quality.htm for a description of our processes and the key performance indicators for the quality of our work.

40. Describe the principles, standards, or guidelines the agency uses as a basis to determine what is considered acceptable quality.

Example answer. See our State's website at www.SNAPInstructions.GenericState.gov/quality.htm for the policies, principles, standards, and guidelines that guide our work.

41. Describe the findings and corrective actions of studies, evaluations or audits to assess compliance with quality standards.

Example answer. Copies of routine inspections results from the state comptroller are posted at www.SNAPInstructions.GenericState.gov/quality.htm Attached is an evaluation of our processes for the 2009 Malcolm Baldrige National Quality Award.

III. Repeated Acquisition Phase

Questions in the Repeated Acquisition phase cover periodic receipt of data. (**11 questions**, including 9 questions repeated from Initial Acquisition)

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.

[11.] Describe each field on the data file and note their valid values. (See template for data dictionary).
Example answer. A data dictionary is attached.

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, reference period, and the target population.

[21.] Were there changes made during the file extract period (e.g., January 2006 through December 2008) or differences across the geographical areas covered by the file in the questions asked that would cause a break in consistency, such as different questions, revised questions, questions in different languages, or deleted questions?

Example answer. A fingerprint status question was added in 2006. Files for benefit months earlier than 2006 do not have this field.

[22.] Were changes made during the file extract period to how the data are processed, such as changes to mode of data collection, changes to instructions for completing the application form, changes to the edit, changes to classification codes, or changes to the query system used to retrieve the data?

Example answer. In January 2007, a new database system was brought online to support the program and pre 2007 data was ported to the new system. To our knowledge, this transition caused no degradation to the data quality.

[23.] Were there changes during the file extract period to geographical boundaries?

Example answer. Coding for geography was refined in June 2006 with the inclusion of a ZIP + 4 address field.

[24.] Were there changes during the file extract period or differences across geographical areas covered by the file that influenced who participated in the program, such as legislative changes, changes to eligibility requirements, expansions of the program, or natural disasters impacting program participation?

Example answer. In August 2006, tornadoes destroyed several large businesses in one county leading to unemployment/reduced income in that area which increased the number of participants meeting program participation requirements.

[25.] Were there any other changes during the file extract period or differences across the geographical areas covered by the file that would cause a break in consistency in regards to comparing data over time? If so, describe the change/difference and when/where it occurred.

Example answer. None

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.

[28.] What is the proportion of duplicate records on the data file?

Example answer. None. The file extracted for benefit data contains a composite primary key of Case ID-Benefit Month-Client ID that ensures one unique occurrence for each combination of these fields. There are regular quality control processes that prevent a single client from being assigned to multiple client ID's and a single Client ID's from being assigned to multiple Case ID's for a given benefit month .

[29.] What is the proportion of missing values for each field? (Feel free to provide an attachment.)

Example answer. Ethnic group is the primary field with missing values. Item nonresponse for the ethnic group field in fiscal year 2005-2010 was less than 1%.

[30.] Describe other metadata provided with the administrative data file such as record counts, range of values, and frequencies of responses.

Example answer. Metadata will include record counts and a data dictionary. The frequency of missing values for each field in the fiscal 2004 data is attached.

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.

42. Describe corrective actions taken to improve the quality of your processes and data.

Example answer. None

43. For the file extract period, describe any new records or revisions to existing records that may occur after data acquisition.

Example answer. Records do get added to the system for a particular month after the end of the month passes, but the lag is at most a day or two. Therefore, the file extract we send you will contain all the records we have and will ever have for the month. However, various fields such as income are subject to revision as new information comes to the SNAP office. After you receive the data file, there could be revisions to records for these fields that you would not have in the file extract we send you.

Example Data Dictionary for Repeated Acquisition Phase

File Name:	Program Beneficiaries.por	Description:	2011 Beneficiary Contact Information
File Date:	January 12, 2012	File Format:	SAS portable dataset

Field Name	Description	Req.	Type	Len	Format	Units	Valid Values	Definitions	Notes:
case	Case ID	Y	Num	6			1 - 999999	Sequential number assigned to each case	Primary Key
yrmo	First benefit month case appeared at address	Y	Num	4	yymm		04 01 - 12	yy – two digit year mm – two digit month	
resaddr	Residential address	Y	Char	1			Null Y N	No response **(Default)** Yes this is a residential address No this is not a residential address	
HHName	Name for household	Y	Char	40			Alphabetic		
addrln1	Mailing address – line 1	Y	Char	22			Alpha-numeric		
addrln2	Mailing address – line 2	N	Char	22			Null Alpha-numeric	No response **(Default)**	
city	City	N	Char	16			Alphabetic		
St	State	N	Char	2			Alphabetic	See Notes	Definitions are contained the USPS standard for ZIP codes dated Jan 1, 2005
Zip	5 digit zip code	Y	Num	4			See Notes	See Notes	Valid values and definitions are contained in USPS standard for ZIP codes dated Jan 1, 2005
Zip4	Additional 4 numbers to 5 digit zip code	N	Char	4			See Notes	See Notes	Valid values and definitions are contained in USPS standard for ZIP codes dated Jan 1, 2005
fips_county	FIPS County code	N	Num	3			See Notes	See Notes	Valid values and definitions are contained in FIPS standard dated Oct 24, 1998.

Example Data Dictionary for Repeated Acquisition Phase

File Name:	Benefits_2011.por	Description:	Benefits data for 2011.
File Date:	January 12, 2012	File Format:	SAS portable dataset

Field Name	Description	Req.	Type	Len	Format	Units	Valid Values	Definitions	Notes:
case	Case ID	Y	Num	6			0 - 999999	Sequential number assigned to each case	Primary Key
yrmo	Benefit month	Y	Num	6	yyyymm		2006 01-12	yyyy – 4 digit year mm – 2 digit month	Primary Key
CID	Client ID	Y	Char	6			0 - 999999	Sequential number assigned to each client	Primary Key
client	Name of client	Y	Char	40			Alphabetic	Full name of client	
ssn	Social Security Number	N	Num	9	###-##-####		Null Numeric	**(Default)**	
birthdt	Birthdate	Y	Num	6	mmddyy		01-12 01-31 00-99	mm – 2 digit month dd – 2 digit day yy – 2 digit year	
race	Ethnic Group	N	Char	1			1 2 3 4 5 6	White Black Hispanic American Indian or Alaskan Native Asian or Pacific Islander Omitted **(Default)**	
sex	Client gender	N	Char	1			M F U	Male Female Unknown **(Default)**	
sig1	Code which identifies the client's status in group	Y	Char	1			A B C D E F G H J K M	Head of Household member Student Striker Military member Eligible legalized alien Treatment center residents Head of Household - nonmember Refugee Disqualified as ineligible alien Disqualified failed to meet SSN requirement	

Example Data Dictionary for Repeated Acquisition Phase

Field Name	Description	Req.	Type	Len	Format	Units	Valid Values	Definitions	Notes:
							R S T U X W Z	Migrant, out of work stream DHS employee Alien with acceptable alien status Disqualified for intentional program violation Seasonal farm worker Non primary wage disqualified for noncompliance Migrant, in work stream Alien granted temporary resident status	
educat	Client Education Level	Y	Char	1			1 2 3 4 5 6 7 8 9 A B C D E F N	First Grade Second Grade Third Grade Fourth Grade Fifth Grade Sixth Grade Seventh Grade Eighth Grade Ninth Grade Tenth Grade Eleventh Grade High School Graduate/ GED Currently Attends JR, High, GED Attending or completed some college Graduate of a 4 yr college No Formal Education	
fingprcd	Finger print status	Y	Char	1			A B C D E F I	Client has filed an appeal and is continuing to receive benefits while appeal is in progress Religious exemption Certified out of office Physically unable to have fingerprints taken Equipment failure Disqualified from the program Fingerprint images have been taken. Client has since been denied. Client record is in inactive file.	

Example Data Dictionary for Repeated Acquisition Phase

Field Name	Description	Req.	Type	Len	Format	Units	Valid Values	Definitions	Notes:
							Y Z	Fingerprint images have been taken. Client currently receiving benefits. Client record is in active file Fingerprint images of one finger has been taken. The image of the other finger is still required. Client record may be in either active or inactive file.	
casestat	Case status code	Y	Num	3			1 2 3 4	Active Denied Suspended Hold **(Default)**	
coupon	Benefit amount	Y	Num	3	$###	Dollar	0 - 999	$0 (Min) **(Default)** $999 (Max)	
serlvl	The client's time limited benefits tier level	Y	Char	1			1 2 3 4 5 6 7 8	Likely to obtain above min. wage within 6 mo. Can obtain above min. wage employment within 1 yr with case management Client has severe impediment to employment Client refused to cooperate Client has 18 months of work experience or completion of HS, GED, Post Sec., Tech or Voc. School Client has 6-17 months recent work experience or completion of 11th grade Client has < 6 months of work experience and completion of less than 3 years of Plucker assigned default for SIG 7,8 clients	

Appendix A—Compendium of Questions Across All Three Phases

No.	Phase of MOU process			Questions, by Dimension
	Discovery	Initial Acquisition	Repeated Acquisition	
Relevance. Relevance refers to how well the administrative data file 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.				
1	✓			What are the general contents of the of the administrative data file(s) maintained by the program?
2	✓			What entity is represented at the record level on the administrative data file (person, household, family, business, etc.)?
3	✓			Describe the scope of the records included on the administrative data file (program participants, geographic areas, etc.).
4	✓			What are the earliest and latest dates for which administrative data are available (including in archive)?
5	✓			How often are the data collected at the program office? (Daily for flow data, or a particular date such as April 1)?
6	✓			What is "reference time period" by which the data are organized in the administrative file(s)?
7	✓			How soon after a reference period ends can a data file extract be prepared and provided?
Accessibility. Accessibility refers to the ease with which the data file extract can be obtained from the administrative agency. This includes the suitability of the form or medium for transferring the data, confidentiality constraints, and cost.				
8	✓			Describe any legal, regulatory or administrative restrictions on access to the data file extract.
9	✓			Describe any records or fields that were collected but cannot be included (for confidentiality or other reasons) in the data file extract provided to the user.
10	✓			What is the typical reimbursement for sharing your data? Are reports, a data product, or a monetary fee given in return?
11		✓		Describe any disclosure control methods or other modifications of the data file used to protect the confidentiality of the data (i.e., Topcoding, Replacing names with identification numbers) or will non-modified data from the program agency's database(s) be provided and accessible to the statistical agency?.
12		✓		How will the data file extract be transferred to the user?
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.				
13	✓		✓	Describe each variable on the administrative data file and note their valid values. (See template for data dictionary).
14	✓			If a complete data dictionary is not available, describe the **primary** identification and response fields on the administrative data file and note their valid values.

Appendix A—Compendium of Questions Across All Three Phases

No.	Phase of MOU process			Questions, by Dimension
	Discovery	Initial Acquisition	Repeated Acquisition	
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. (continued)				
15		✓		If a complete data dictionary is not available, describe any **remaining** fields (not already documented in the Discovery phase).
16		✓		Provide a copy of any forms and instructions to the applicant used to collect data.
17		✓		Describe the methodology used to recode original data to create a value for a new variable, such as assigning a reported age to an age range.
18		✓		Describe the data architecture and relationship 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?
19		✓		Does the dataset contain records for those who are denied eligibility to the program or benefits and how can those records be identified?
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 file extract period, and the target population.				
20		✓		Please describe any classification systems used for categorizing or classifying the data. For example: Do questions asked about race use the current race and ethnicity categories defined by the Office of Management and Budget (OMB)? Do questions asked about industry classifications use the current North American Industry Classification System (NAICS)?
21		✓	✓	Were there changes made during file extract period (e.g., January 2006 through December 2008) or differences across the geographical areas covered by the file in the questions asked that would cause a break in consistency, such as different questions, revised questions, questions in different languages, or deleted questions?
22		✓	✓	Were changes made during the file extract period to how the data are processed, such as changes to mode of data collection, changes to instructions for completing the application form, changes to the edit, changes to classification codes, or changes to the query system used to retrieve the data?
23		✓	✓	Were there changes during the file extract period to geographical boundaries?
24		✓	✓	Briefly, describe substantial changes during the file extract period or differences across the geographical areas covered by the file that influenced who participated in the program, such as legislative changes, changes to eligibility requirements, expansions of the program, or natural disasters impacting program participation?
25		✓	✓	Were there any other changes during the file extract period or differences across the geographical areas covered by the file that would cause a break in consistency in regards to comparing data over time? If so, describe the change/difference and when/where it occurred.

Appendix A—Compendium of Questions Across All Three Phases

No.	Phase of MOU process			Questions, by Dimension
	Discovery	Initial Acquisition	Repeated Acquisition	
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.				
26		✓		What investigations/analyses have been conducted that reveal data quality characteristics (such as GAO reports, Office of Inspector General audits, internal agency reviews, etc.)?
27		✓		What percentage of eligibles are not included on the data file or what percentage of those mandated are not compliant? What is known about their characteristics?
28		✓	✓	What is the proportion of duplicate records on the data file extract?
29		✓	✓	What is the proportion of missing values for each field? (Feel free to provide an attachment.)
30		✓	✓	Describe other metadata provided with the administrative data file such as record counts, range of values, and frequencies of responses.
31		✓		What are the known sources of errors in the administrative data (e.g. non-response, keying, coding errors)?
32		✓		What questions are most often misinterpreted?
33		✓		Which items are subject to revision either by editing or updating data values? What items are revised the most?
34		✓		If a value is revised from what was originally reported, describe any flags in the dataset that would indicate that the value was changed as well as explain why the value was changed.
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.				
35		✓		Describe the purpose of the administrative program.
36		✓		Describe your processes for data collection, editing, review, correction, dissemination, and retention.
37		✓		Who is source data collected from (self-reported, interview a third party)?
38		✓		How is source data collected (paper questionnaire, computer assisted person interview, computer assisted telephone interview, web data collection form)?
39		✓		Describe quality control checks and the typical results for your production processes.
40		✓		Describe the agency's quality standards.
41		✓		Describe the findings and corrective actions of studies, evaluations or audits to assess your compliance with quality standards.
42			✓	Describe corrective actions taken to improve the quality of your processes and data.
43			✓	For the file extract period, describe any new records or revisions to existing records that may occur after data acquisition.

Appendix B—Data Dictionary Template

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