REPORT OVERVIEW

In the Fall of 2010, the Bureau of the Census, Geography Division contracted with independent subject matter experts David Cowen, Ph.D., Michael Dobson, Ph.D., and Stephen Guptill, Ph.D. to research five topics relevant to planning for its proposed Geographic Support System (GSS) Initiative; an integrated program of improved address coverage, continual spatial feature updates, and enhanced quality assessment and measurement. One report frequently references others in an effort to avoid duplication. Taken together, the reports provide a more complete body of knowledge. The five reports are:

1. Reporting on the Use of Handheld Computers and the Display/Capture of Geospatial Data
2. Measuring Data Quality
3. Reporting the State and Anticipated Future Directions of Addresses and Addressing
4. Identifying the Current State and Anticipated Future Direction of Potentially Useful Developing Technologies
5. Researching Address and Spatial Data Digital Exchange and Data Integration

The reports cite information provided by Geography Division staff at "The GSS Initiative Offsite, January 19-21, 2010." The GSS Initiative Offsite was attended by senior Geography Division staff (Division Chief, Assistant Division Chiefs, & Branch Chiefs) to prepare for the GSS Initiative through sharing information on current procedures, discussing Initiative goals, and identifying Initiative priority areas. Materials from the Offsite remain unpublished and are not available for dissemination.

The views expressed in these reports are the personal views of the authors and do not reflect the views of the Department of Commerce or the Bureau of the Census.

Department of Commerce
Bureau of the Census
Geography Division

Task 5: Measuring Data Quality

A Report to the Geography Division, U.S. Bureau of the Census
Deliverables #11 & 12 - Syneren Technologies Contract – Task T005

Stephen Guptill, Ph.D., Lead Author,
Michael Dobson, Ph.D., and David Cowen, Ph.D.

December 29, 2010

Executive Summary:

The success of the GSS Initiative hinges on being able to measure data quality. As stated in the GSS Initiative draft operational plan, without data quality evaluation and reporting procedures, the Census Bureau will not have suitable evidence that the MAF/TIGER data base (MTdb) of addresses and geographic features will meet acceptable quality standards for program use. In addition, the Census Bureau will not be able to adequately determine whether a targeted approach to the 2020 address canvassing operation is feasible. This report lays out a framework to establish a program for measuring data quality.

If possible, the data quality program should build upon current best practices within the discipline. A variety of federal, industry and international standards exist for the specification, evaluation, and reporting of spatial data quality information. The 19100 Geographic Information Standard Series developed by the International Organization for Standardization (ISO) is most applicable for use by the Geography Division. Utilizing the Geography Division's data quality requirements, we specified ISO 19100 compliant sets of data quality elements, sub elements, measures, and evaluation procedures. As part of its data quality management program, Census should adopt and utilize the language, structures, and information infrastructure for data quality evaluations as promulgated by the ISO 19100 suite of standards. The Federal Geographic Data Committee (FGDC), in September 2010, adopted more than 25 items from the ISO 19100 suite, making the Census Bureau responsible for compliance where applicable.

However, before data quality evaluations can begin, a rigorous set of data specifications are required. Within the Geography Division some data definitions exist, but they lack the level of specificity and precision necessary to perform internal data quality testing and external data exchange. The Geography Division must review and revise as necessary the data content specifications for the MTdb so that the elements are precisely and formally defined in terms of the requirements that they are intended to fulfill.

To begin implementation of a data quality management program, the Geography Division needs to validate and revise, as necessary, the set of data quality elements proposed in this report against Census user requirements. In addition, the Geography Division needs to: establish appropriate conformance levels for each data quality measure, choose what data quality evaluation procedures will be employed, and determine when in the data product life cycle they will be applied. When implemented, these processes will capture information not only about new collections, but also provide an estimate of the quality of the current MTdb.

By rigorously recording the process history of the contents of the MTdb, the Geography Division has captured an unknown, but perhaps significant amount, of nascent data quality information. In addition, the 2010 Census: Revised Quality Control Plan for the Address Canvassing Operation specified a set of quality measurement procedures for much of the MTdb content, but the disposition of that data is unclear. The Geography Division should establish an initial set of data quality values for the MTdb by: processing and analyzing the existing metadata holdings and placing the appropriate contents into ISO 19100 compliant form, and obtaining and analyzing the results from the Quality Control Plan for the Address Canvassing Operation to create a baseline measure of MTdb data quality.

Maintenance of the MTdb is not a "closed" production process. Multiple data sources, most of which are outside of the Geography Division's direct control, are used in updating the MTdb. Data quality principles (many of which are now instantiated in the Geography Division's Business Rules Engine) must be incorporated into all of the MTdb data flows. In particular, the Geography Division should utilize data specifications, data quality measures, and conformance levels in evaluating potential data exchanges/acquisitions with third parties. Data evaluation software should be utilized as part of the data production process, particularly in handling contributions from data partners. Additionally, the Geography Division should refine and continue to apply data quality procedures and processes to field update operations.

Other Census groups use MTdb data for various purposes and these groups have their own quality assurance procedures. When MTdb data fail the user's data quality requirements, a feedback loop should be established in which the users (e.g. other Census Bureau Divisions) return deficient data to the Geography Division for analysis. In essence, the Geography Division should be made aware of what data (elements) failed; test those elements against the data quality metrics to see where the metrics failed; and determine what corrective action is needed.

As geographic information permeates more aspects of daily life, data users' requirements for knowledge about quality in geographic information continue to increase. This has spurred commercial software vendors to devote more resources towards data quality standardization efforts and to develop software to perform testing and reporting in conformance with ISO 19100 standards. Examples of such software include ArcGIS DataReviewer and 1Spatial Radius Studio. The Geography Division should evaluate this type of software for its applicability to Census operations.

Defining and implementing a comprehensive data quality management program for the MTdb is a difficult task. But it is a challenge that the Geography Division, based on its past record of innovation and success in this field, can meet.

Scope and Purpose:

In organizing the planned activities contained in the Geographic Support System (GSS) Initiative, it is useful to think of the continual update of the Master Address File/Topologically Integrated Geographic Encoding and Referencing database (MTdb) as a cyclical process. First, data requirements are identified and defined in detail. Then the data maintenance resources of the Census staff, external partners, and data suppliers are utilized as appropriate. New technology and processes are incorporated to meet data collection needs, perform quality control on data, and improve operations. Most importantly, quality assurance methods are constantly employed to ensure that the data are authoritative and that the quality of the data are maintained. At any step in this process, new data needs, partners, or technology may be identified or discovered, so the cycle repeats, incorporating these refinements. Under this paradigm, the activities identified in the operational plan can be identified, integrated, sequenced, and measured providing a cycle of continuous improvement.

Understanding data quality is key to the success of the entire quality improvement process. In the cycle of continually maintaining the MTdb, one must be able to affirm that new data added to the MTdb meet the criteria for inclusion, and that any revision or replacement data are of better quality than the existing holdings. These quality issues are tightly coupled with trying to answer questions such as: “why can’t we geocode this address?”, “where do we need to do address canvassing?”, “can we locate every housing unit?”, or “what is new or different in this area?” Answering these questions through the implementation of data quality measures is often quite complex in practice.

The measuring data quality task addresses three major requirements:

- Identify and describe the data quality elements that are associated with the address and spatial components of the MTdb. Ideally these studies would deal with the entire content of the MTdb, but measuring the data quality of every feature would have high costs and perhaps limited utility. **Thus, we will concentrate our efforts on three classes of data that likely will yield high returns for the data quality investment: addresses, the road network, and the MAF Structure Points (MSPs).**

- Develop specific methods for measuring and reporting the quality of applicable data elements (i.e. data quality procedures) while utilizing best professional practices and employing national, international, and industry standards.

- Identify and research relevant components for the successful exchange of address and spatial data between the Census Bureau and their partners.

The results of these investigations are reported and are used to relate data quality elements, data quality procedures, and data reporting practices to applicable Census update operations.

Adherence to national, international, and industry standards is important to the Census Bureau, and we have been tasked to ensure that our research reflects this body of work. Thus, this report

begins with a discussion of the activities of various standards organizations and their work with respect to spatial data quality.

Standards Organizations and Spatial Data Quality:

Three organizations have the most relevance with respect to standards related to the geographic databases of the Census Bureau. These are the International Organization for Standardization (ISO) Technical Committee 211 on Geographic Information/Geomatics (TC 211), the industry led consortium the Open Geospatial Consortium (OGC), and the Federal Geographic Data Committee (FGDC).

The standardization efforts related to digital geographic data began in the early 1980's. National organizations such as the FGDC and international professional societies, e.g. the International Cartographic Association (ICA) were developing standards for the transfer and exchange of geographic data between computer systems. At the time, there were no standards that had broad international support. By 1995, ISO/TC 211 began developing international standards for spatial data and the OGC was developing computer interface specifications for international adoption. Subsequently ISO/TC 211 and the OGC formed a joint coordination group to leverage mutual development and minimize technical overlap. The OGC submits their specifications for ISO standardization via ISO/TC 211.

At the present time ISO/TC 211 has developed over 45 standards related to geoinformation and geomatics. Of these, the following have the most relevance to this study of spatial data quality:

- ISO 19113:2002 Geographic information — Quality principles
- ISO 19114:2003 Geographic information — Quality evaluation procedures
- ISO 19115:2003 Geographic information — Metadata
- ISO 19131:2007 Geographic information — Data product specifications
- ISO/TS 19138:2006 Geographic information — Data quality measures
- ISO/CD 19157 Geographic information — Data Quality

It should be noted that TC211 has a work item, 19157, the scope of which is to merge and revise 19113, 19114 and 19138. ISO 19157 "Geographic Information – Data Quality" is in a draft form and has been submitted to members for ballot and comments. When approved, perhaps in 2011, ISO 19157 will cancel and replace ISO 19113, ISO 19114 and ISO/TS 19138. ISO 19115 is also undergoing a systematic review (scheduled to be completed in 2012) which should allow the working teams to ensure that existing discrepancies between the standards will be eliminated. Details on the standards activities of TC211 are described in their *Standards Guide* (ISO/TC211, 2009) available at http://www.isotc211.org.

The OGC oversees the development of a set of standards and specifications for use by the GIS industry. OpenGIS® Standards and Specifications are technical documents that detail interfaces or encodings. Software developers use these documents to build support for the interfaces or encodings into their products and services. These specifications are the main "products" of the OGC and have been developed by the membership to address specific interoperability challenges. The OpenGIS® Web Feature Service Implementation Specification and the Web Map Service Implementation Specification are examples of OGC specifications that could be of

use to the Geography Division. As the ISO 19100 standards are implemented by the OGC and its industry members, the Census Bureau can take advantage of commercial off-the-shelf (COTS) software that automates the recording of specified data elements and generates reports and data product packages that adhere to the ISO standards.

The FGDC is a United States federal government interagency committee that promotes the coordinated development, use, sharing, and dissemination of geospatial data on a national basis. The FGDC develops geospatial data standards for implementing the National Spatial Data Infrastructure**,** in consultation and cooperation with State, local, and tribal governments, the private sector and academic community, and, to the extent feasible, the international community. The FGDC develops geospatial data standards only when no equivalent voluntary consensus standards exist, in accordance with Office of Management and Budget (OMB) Circular A-119. An important role of the FGDC in the standards arena is that, unlike ISO and OGC, it will specify mandatory and voluntary content guidelines for use by all parties. For example, the FGDC Metadata Standard (*Content Standard for Digital Geospatial Metadata* (CSDGM), Ver. 2 (FGDC-STD-001-1998)) must be used by Federal agencies to document their geospatial data holdings. This standard will be superseded, eventually, by the North American Profile (NAP) of ISO 19115, Geographic information. Upon adoption and sufficient time for vendors to create robust tools to create NAP metadata and for CSDGM Metadata to be converted to the NAP, then the CSDGM will be replaced by the NAP. This is likely to occur in the 2011-12 timeframe.

On September 27, 2010, the FGDC endorsed 64 non-Federally authored geographic information standards and specifications from ISO, OGC, and other organizations (the list can be found at: http://www.fgdc.gov/standards/fgdc-endorsed-external-standards/index_html). In this context, endorsement means that the non-Federally authored standard or specification has the same status as that of an FGDC developed standard, i.e., its use by Federal agencies is mandatory. The Geography Division needs to be aware of the utility and applicability of these standards and specifications as it implements the GSS Initiative.

Importantly, in 2010, the FGDC released for public comment the draft *United States Thoroughfare, Landmark, and Postal Address Data Standard*. This standard covers data content, data classification, data exchange, and data quality. The Urban and Regional Information Systems Association (URISA), in conjunction with the FGDC Subcommittee on Cultural and Demographic Statistics chaired by the U.S. Census Bureau, developed the draft standard.

The *United States Thoroughfare, Landmark, and Postal Address Data Standard* has been created to:

- Provide one standard that meets the diverse address data management requirements for local address administration, postal and package delivery, emergency response (and navigation generally), administrative recordkeeping, and address data aggregation.
- Support the use of best practices in address data management.
- Provide a systematic, consistent basis for recording all addresses in the United States.
- Define the elements needed to compose addresses and store them within relational databases and geographic information systems.

- Define the attributes needed for address documentation, mapping, and quality testing, including address ID's, coordinates, and linear reference locations.
- Provide a complete taxonomy (systematic classification) of US addresses that is useful to address data managers.
- Introduce the idea of the address reference system—the formal description of the local address assignment rules, both spatial and non-spatial—and define its elements and attributes, as a basis for address assignment and quality testing.
- Define tests and procedures for address data quality testing, error-trapping, and anomaly identification.
- Support seamless exchange of address information, and foster consistent implementation of this standard, by defining XML models for every address element, attribute, and class, integrated into a single XML Schema Document.
- Offer a migration path from legacy formats to standards-compliant ones.
  (Source: FGDC, 2010, p. 4)

The standards and guidelines developed by these organizations provide a useful (and sometimes mandated) framework for the developments that support the Geography Division's operational missions. The Census Bureau benefits by adopting standardized methods to describe and communicate data quality information among disparate producers, users, and information systems. One challenge facing the Census Bureau is to identify the specific data quality elements, measures, and procedures that will enable it to pursue the programs of the GSS Initiative.

Elements of Spatial Data Quality

This section provides a brief overview of the major elements of spatial data quality. Several books on this topic have been published, most notably: *Elements of Spatial Data Quality* (Guptill and Morrison, 1995), *Accuracy of Spatial Databases* (Goodchild and Gopal, 1989), *Data Quality in Geographic Information, From Error to Uncertainty* (Jensoulin and Goodchild, 1998), and *Spatial Data Quality* (Wenzhong, Fisher, and Goodchild, 2002). The concepts and principles established in these works formed the foundation for the content of the various ISO standards on spatial data quality. In order to establish a common set of terminology in discussing these topics, we have chosen to utilize the language of ISO TC211 and its terms and definitions.

The major rationale for characterizing spatial data quality is noted by the ISO TC211 Secretariat:

> *The purpose of describing the quality of geographic data is to facilitate the selection of the geographic dataset best suited to application needs or requirements. Complete descriptions of the quality of a dataset will encourage the sharing, interchange and use of appropriate geographic datasets. A geographic dataset can be viewed as a commodity or product. Information on the quality of geographic data allows a data producer or vendor to validate how well a dataset meets the criteria set forth in its product specification and assists a data user in determining a product's ability to satisfy the requirements for their particular application.*

> *This International Standard [ISO 19113] establishes the principles for describing the quality of geographic data and specifies components for reporting quality information. It also provides an approach to organizing information about data quality. This International Standard does not attempt to define a minimum acceptable level of quality for geographic data.* (Source: ISO TC 211, 2009, p.50)

The information on data quality (which is encapsulated in metadata records in accordance with ISO 19115) can be applied at varying data granularities, from individual feature instances to entire databases. A quality description can be applied to a dataset series, a dataset, or a smaller grouping of data located physically within the dataset sharing common characteristics so that its quality can be evaluated. It is possible, for example, to associate information with an individual address record, noting data such as: when it was changed, the source of the change, if the address is considered deliverable by USPS, and when it was last verified.

The quality of a dataset is described using the following components: data quality elements, data quality sub-elements, and data quality overview elements (Source: ISO, 2002, pp. 5-8).

Data quality elements and sub-elements

Data quality elements and sub-elements, where applicable, are used to describe aspects of the quantitative quality of a dataset:

- completeness – presence and absence of features, their attributes and relationships;
    - commission – excess data present in a dataset,
    - omission – data absent from a dataset.
- logical consistency – degree of adherence to logical rules of data structure, attribution and relationships (data structure can be conceptual, logical or physical);
    - conceptual consistency – adherence to rules of the conceptual schema,
    - domain consistency – adherence of values to the value domains,
    - format consistency – degree to which data is stored in accordance with the physical structure of the dataset,
    - topological consistency – correctness of the explicitly encoded topological characteristics of a dataset.
- positional accuracy – accuracy of the position of features;
    - absolute or external accuracy – closeness of reported coordinate values to values accepted as or being true,
    - relative or internal accuracy – closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true,
    - gridded data position accuracy – closeness of gridded data position values to values accepted as or being true.

- temporal accuracy – accuracy of the temporal attributes and temporal relationships of features;
    - accuracy of a time measurement – correctness of the temporal references of an item (reporting of error in time measurement),
    - temporal consistency – correctness of ordered events or sequences, if reported,
    - temporal validity – validity of data with respect to time.
- thematic accuracy – accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships;
    - classification correctness – comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference dataset),
    - non-quantitative attribute correctness – correctness of non-quantitative attributes (e.g. correctness of attribute values such as “road name” or “pavement type”),
    - quantitative attribute accuracy – accuracy of quantitative attributes.

Additional data quality elements may be created to describe a component of the quantitative quality of a dataset not addressed in ISO 19113.

Data quality overview elements

The data quality overview elements are used to describe the **non-quantitative** quality of a dataset:

- purpose – describes the rationale for creating a dataset and contains information about its intended use;
- usage – describes the application(s) for which a dataset has been used. It describes the uses of the dataset by the data producer or by other, distinct, data users;
- lineage – describes the history of a dataset and, in as much as is known, recounts the life cycle of a dataset from collection and acquisition through compilation and derivation to its current form.

Lineage may contain two unique components:

- source information describing the parentage of a dataset;
- process step or history information describing a record of events or transformations in the life of a dataset, including the process used to maintain the dataset (whether continuous or periodic), and the lead time.

Using data quality elements

These standards can be used when:

- identifying and reporting quality information,
- evaluating the quality of a dataset,
- developing product specifications and user requirements, and
- specifying application schemas.

The quality of a dataset is described using a combination of data quality elements and data quality overview elements. Data quality elements, together with data quality sub-elements and the descriptors of data quality sub-elements, describe how well a dataset meets the criteria set forth in its product specification and provide quantitative quality information about the dataset. Data quality overview elements are critical for assessing the quality of a dataset for a particular application that differs from the intended implementation. Theoretically, a dataset (from the data quality perspective) may be as small as a single feature or feature attribute contained within a larger dataset. When data quality measures are applied to collections or aggregates of individual features, those measures tend to serve as screens or filters. In other words, for example, all of the features in this collection have a positional accuracy greater than 3.5m, or all of the features have passed a logical consistency check.

The ISO standards recognize that quantitative and non-quantitative quality information may have associated quality and the analysis of the quality about quality information may include a measure of the confidence or the reliability of the quality information. This type of information is then recorded in a quality evaluation report. The relationships between these various elements of spatial data quality are shown in Figure 1.

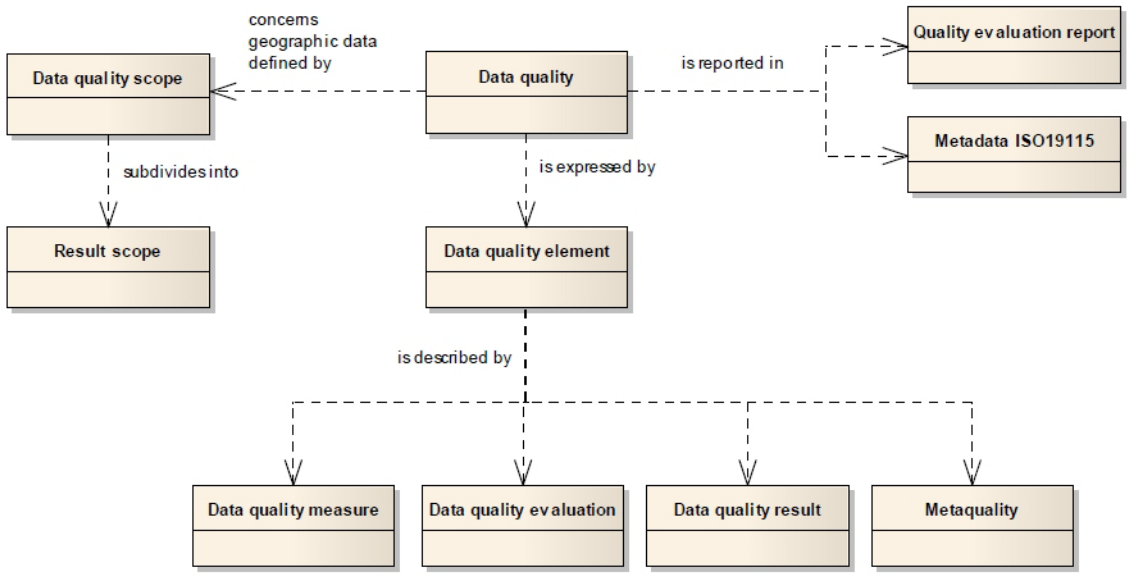

Figure 1 An Overview of Data Quality Elements (Source: ISO, 2010, p. 5)

A register of data quality measures is given in Annex C of the ISO 19157. It contains a set of commonly used data quality measures with all required components for data quality measures as specified in the standard. The register contains multiple measures for each data quality element and the choice of which one to use will depend on the type of the data and its intended purpose. Measures from this register should be used in implementing the standard.

The draft *United States Thoroughfare, Landmark, and Postal Address Data Standard* provides more detailed specifications on data quality, assessment, accuracy, and reporting standards for address data. The Address Data Quality part of the standard specifies a complete suite of data quality tests for all address elements, attributes, and classes.

The standards document and included tests utilize terms consistent with the FGDC's *Content Standard for Digital Geospatial Metadata* and subsequent ISO standards of spatial data quality. As such, the data quality elements used are the same as for spatial data, that is: Attribute (Thematic) Accuracy; Logical Consistency; Temporal Accuracy; Completeness; Positional Accuracy; and Lineage. It is assumed that the terminology changes that result from the migrations from CSDGM to NAP will be implemented in the Address Standard.

The address data quality tests measure how well a given set of address records conforms to the standard and the local address reference system. Each test specification includes the scope, measure, and procedure of the test; an SQL pseudo code script; and parameters for calculating anomalies as a percentage of the data set. The list of these quality measures and tests is given in Table 1. Which, if any, of these measures (and tests) are appropriate within the Census context will be discussed in a later section of this report.

Quality Measure	Quality Element
Address Completeness Measure	Completeness
Address Elevation Measure	Attribute (Thematic) Accuracy
Address Left Right Measure	Logical Consistency
Address Lifecycle Status Date Consistency Measure	Temporal Accuracy/Logical Consistency
Address Number Fishbones Measure	Logical Consistency
Address Number Parity Measure	Logical Consistency
Address Number Range Completeness Measure	Logical Consistency
Address Number Range Parity Consistency Measure	Logical Consistency
Address Number Range Sequence Measure	Logical Consistency
Address Range Directionality Measure	Logical Consistency
Address Ref. System Axes Point Of Beginning Measure	Logical Consistency
Address Ref. System Rules Measure	Logical Consistency
Check Attached Pairs Measure	Logical Consistency
Complete Street Name Tabular Domain Measure	Attribute (Thematic) Accuracy
Complex Element Sequence Number Measure	Attribute (Thematic) Accuracy
Data Type Measure	Logical Consistency
Delivery Address Type Subaddress Measure	Logical Consistency
Duplicate Street Name Measure	Logical Consistency
Element Sequence Number Measure	Attribute (Thematic) Accuracy
Future Date Measure	Temporal Accuracy
Intersection Validity Measure	Logical Consistency
Left Right Odd Even Parity Measure	Logical Consistency
Location Description Field Check Measure	Positional Accuracy/Lineage
Low High Address Sequence Measure	Logical Consistency
Official Status Address Authority Consistency Measure	Logical Consistency
Overlapping Ranges Measure	Logical Consistency
Pattern Sequence Measure	Logical Consistency
Range Domain Measure	Attribute (Thematic) Accuracy
Related Not Null Measure	Logical Consistency
Related Element Uniqueness Measure	Attribute (Thematic) Accuracy
Repeated Element Uniqueness Measure	Attribute (Thematic) Accuracy

Segment Directionality Consistency Measure	Logical Consistency
Spatial Domain Measure	Positional Accuracy
Start End Date Order Measure	Temporal / Attribute (Thematic) Accuracy
Subaddress Component Order Measure	Attribute (Thematic) Accuracy
Subaddress Element Z Level Measure	Attribute (Thematic) Accuracy
Tabular Domain Measure	Attribute (Thematic) Accuracy
Uniqueness Measure	Attribute (Thematic) Accuracy
USNG Coordinate Spatial Measure	Positional Accuracy
XY Coordinate Completeness Measure	Logical Consistency
XY Coordinate Spatial Measure	Positional Accuracy

Table 1 Quality Measures and Quality Element Type (Source: FGCD, 2010, Sec. 2, Address data content)

According to the draft standard, any packages of address data to be exchanged must include:

- FGDC Metadata, conforming to the FGDC-STD-001-1998 Content Standard for Digital Geospatial Metadata, Version 2.0
- Address data, expressed as an XML document conforming to the Address Standard XML Schema.

However, it should be noted that the provision of any data quality information is optional.

Data quality in the Census context

The set of existing and proposed standards on data quality allow the data producer almost limitless choices for associating quality information with datasets. The real difficulty lies in selecting which database components require data quality measures, at which level of aggregation should these measures be applied (each feature, a group of features, the entire database), and how implementing these measures impact Census workflows and operations.

For example, the positional accuracy of road features (with an acceptance threshold of 7.6 m Circular Error (CE) 95) was determined by comparison of the location of precisely located control points (within a given geographic area) at road intersections to the comparable locations in the MTdb.[1] The Geography Division plans to continue building the inventory of GPS control points to assess the positional accuracy of streets in a minimum of one-eighth of US counties per fiscal year to determine whether updates are necessary to support collecting and assigning census survey data to correct geographic locations. However, no attempt is made to measure the positional accuracy of the curvilinear road segments between intersections. Does this matter? Is it worth the cost to do so? Can conflation routines be used to automatically align roads with high resolution imagery?

[1] In general, the given geographic area was a county and 110 control points were used per county. In some cases less then 110 points were acceptable (there was a sliding scale of points depending on number of existing road intersections in a county; 351+ intersections required 110 points, 201-350 intersections required 88 points, 151-200 intersections required 77 points, 101-150 intersections required 66 points, 76-100 intersections required 55 points, 61-75 intersections required 50 points, 46-60 intersections required 44 points, 0-45 intersections required 40 points). It is important point to note that control points could only be collected at intersections that already existed in TIGER. (Source: L. Godwin, Census Bureau, personal communication)

The TIGER database contains hundreds of feature types and millions of feature instances. An individual data quality measure could be attached to every feature type and each feature instance. However, it is hard to imagine the practical value of doing so, particularly when some features (for example towers) are of less importance to Census operations than others (such as roads). Since each element of data quality would have an associated cost, we need to understand the relative importance of the content of the MTdb to the mission of the Census. Factors to consider in weighing the value of a feature or feature attribute include: the importance of the data element, its part in maintaining topology, and where it fits within a hierarchy of data usage.

Address data has unique data quality issues. As the draft FGDC Address Standard notes:

> *Measuring quality requires a definition for quality. There are definitions of quality in existing standards. Those describing addresses, however, describe the utility of a data set for specific purposes. The National Emergency Number Association (NENA), for example, has documented exchange standards that include a way to describe address quality relative to an established Automatic Location Information (ALI) files. Similarly, the United States Postal Service (USPS) Postal Addressing Standards describe addresses as used for mailing. These application-specific assessments fulfill the purpose of their respective documents.*
>
> *Assessing the quality of address data content independent of formats or specific uses, however, is a very different task. It requires information on each of the many aspects of address information. Evaluations for a specific purpose can be constructed from that information, with criteria varying according to each application. What is needed is a definition of the information required, a definition of the elements of address quality.* (Source: FGDC, 2010, p. 294)

Again, perhaps the most important part of this activity is to determine the data elements that are important to Census operations and how incorporation of data quality measures can enhance the efficiency and effectiveness of those operations. For addresses, data elements that measure factors that effect mailability, deliverability, locatability, and geocodability are paramount.

Review of Existing Data Specifications and Quality Measures:

The first phase of this research is to summarize the data specifications and quality measures currently in use within the Census Geography Division with respect to addresses, roads, and MAF Structure Points (geographic locations of structures corresponding to MAF units).

Data specifications play a major role in quality evaluations. They provide the “reality” against which the data collection is compared. For example, consider testing the road feature set in TIGER for “completeness”. What is the specification for the collection of a “road”? Is there a minimum length criterion? How does a “private driveway” differ from a “private road”? Are only private driveways of a greater than specified length collected? Unless there is an unambiguous data collection specification, you do not have the standard against which you can evaluate completeness.

It is important to understand the data quality measures that the Geography Division has in place within MAF and TIGER. For example, are the Spatial Metadata Identifiers (SMIDs) an adequate measure of the positional accuracy of the road features? Are the processing and history fields embedded in the MAF appropriate records of data lineage? We will discuss the data quality measures currently in place, as determined from the information available to the consulting team.

Addresses

The Master Address File (MAF) contains an inventory of all known living quarters in the United States, Puerto Rico and associated island areas[2]. The MAF is used to support most of the Census Bureau operations including the decennial census, the American Community Survey and other demographic surveys. The content of the MAF includes address information, Census geographic location codes, as well as source and history data. In 2010, the MAF contained over 270 million address records linked to over 190 million MAF units (a MAF Unit may be associated with more than one address record).[3]

The Census Bureau divides all addresses into two classes: residential and nonresidential. Residential addresses are addresses of housing units (HU) and/or group quarters (GQ) where one or more people could live. Nonresidential addresses are addresses of a structure or unit within a structure that do not serve as a residence, such as commercial establishments, schools, government offices, and churches. The MAF contains only a selected subset of nonresidential addresses. Some of the nonresidential addresses that exist in the MAF are addresses that were thought to have been HUs, that were later determined to be nonresidential (through field work or through LUCA). Some structures can contain both residential and nonresidential units, even though they have a single address.

There are two major types of address formats: city-style and noncity-style. The majority (about 90%) of housing units and group quarters in the United States have a city-style address type, with a house number, street name address and ZIP code. For example: 123 Elm St., 13752; 200 Clark Apt. 316, 97210; 101 N Church St., 22180. Addresses that do not include a house number and street name are noncity-style addresses. While majority of noncity-style addresses are located in the more sparsely settled areas of the United States they may exist in more populated locales. The noncity-style mailing addresses include two types: Post Office box, and rural (or highway contract) route and box number. An additional address type, location description, consists of a description of a location or housing unit (e.g. White House Blue Shutters, Corner of Timber Rd and County Rd 15) and is used in conjunction with the other types of addresses in the MAF.[4]

This information is held in a variety of tables within the MTdb, with linkages between them. The MAF is created from a number of sources (such as USPS DSF files, LUCA submissions, AdCan

[2] Island area addresses are not yet included in the MTdb.
[3] MAF statistics presented at GSS Initiative Offsite Meeting. Briefing: *Addresses 101: A Background Primer*, by Shawn Hanks, handout p. 13.
[4] Location descriptions may be collected for all types of addresses - and depending on the requirements of the end customer may be used in lieu of other types of addresses. For example, in U/L many MAF records had PO Boxes but because enumerators were locating the housing units in the field, Location Description information was printed on the listing pages used by these enumerators (Source: A Johnson, Census Bureau, personal communication)

updates, and other census and survey operations) with data provided in differing formats. Table 2 (below) shows the range of activities and sources utilized to create and improve the content of MAF for the 2000 and 2010 decennial censuses. These data inputs are processed through various MAF software routines to standardize the address formatting and to populate the fields in the MAF database. The address elements used in the MTdb are all present in the draft *United States Thoroughfare, Landmark and Postal Data Address Standard.*

After discussions with the Census Bureau staff, we have concluded that address data quality information is largely contained in metadata files that document the lineage of an address or set of addresses. Presumably this includes information such as the initial source of the address, which software processes were utilized for verification and logical checking, if this address is a variant form of an existing address, and how an address might have been changed during field checks or local reviews (shown in Table 2). However, this is speculation on our part since detailed descriptions of the MAF metadata holdings were not available for use in this report.[5]

Census	Original Sources of List(s)	Field Checks of List	Local Review of List
2000	• Master Address File (MAF) City-Style Areas (81% of population): started with 1990 ACF, updated every 6 months with Postal Service Delivery Sequence File (DSF) • MAF Non-City-Style or Update/Leave/Mailback Areas (16% of population): census enumerators built list in 1998–1999 • List/Enumerate Areas (1% of population): no advance list • Special Places List (3% of population): compiled from variety of sources	• MAF City-Style: – Block Canvass (1999, census enumerators rechecked entire list) – Reconcile (verify) LUCA changes (see next column) – Incorporate updates from Postal Service's final intensive check of DSF prior to delivery • MAF Update/Leave: census enumerators updated at time of delivery	• MAF City-Style: – Local Update of Census Addresses Program 1998: local review of MAF; local review of Census Bureau feedback; review of local appeals by OMB – New Construction LUCA (early 2000) • MAF Update/Leave: Local Update of Census Addresses Program 1999—similar to MAF City-Style Areas but later • Special Places (Group Quarters): LUCA review opportunity in early 2000
2010	• Master Address File (MAF) City-Style Areas (76% of population): started with 2000 MAF, updated every 6 months with Postal Service DSF • MAF Non-City-Style or Update/Leave/Mailback Areas (20% of population): started with 2000 MAF, updated every 6 months with Postal Service DSF • List/Enumerate Areas (1% of population): no advance list • Group Quarters List (3% of population): compiled from variety of sources	• MAF City-Style areas: Address Canvass (2009, entire list); included group quarters • MAF Update/Leave areas: – Address Canvass (2009, entire list) – Census enumerators will update at time of delivery • Group Quarters: list validation, Fall 2009	• MAF City-Style: – Local Update of Census Addresses Program: local review of MAF; local review of Census Bureau feedback; review of local appeals by OMB – New Construction LUCA (early 2010) • MAF Update/Leave: Local Update of Census Addresses Program, similar to MAF City-Style Areas • Special Places (Group Quarters): LUCA review opportunity combined with household LUCA

Table 2 Coverage Improvement Programs and Procedures (Source: NRC, 2010, p. 44)

There are two explicit quality related fields in the "Address" table in MAF. These fields flag a record as a "preferred location" and/or a "preferred mailing [address]". As noted in the MAF/TIGER 1.1 Physical Database Design:

> "Each MAFUNIT can have multiple addresses, each flagged as mailing, location, or both. Only one can be the preferred mailing address and only one (possibly the same) can be the preferred location address."

[5] The INSPIRE Address Data Specification (INSPIRE, 2009), offers examples of tracking the lineage (or lifecycle) of an address from proposed to reserved, current, and retired.

There are no explicit fields in MAF that correspond to the data quality measures enumerated in the draft *United States Thoroughfare, Landmark and Postal Data Address Standard* (listed earlier in Table 1). It is quite possible that equivalent data quality information exists in the MAF and could be inferred from the internal processes and rules used to populate, match, certify, and maintain MAF. Recommendations data quality measures to be considered for explicit valuation are presented later in this report.

Road Network

In the United States there are no agreed upon standards for definitions of the features that comprise the road network. Likewise clear, unambiguous, definitions for attributes, attribute values, delineation rules, data collection (extraction) rules, composition rules, and delimiting rules do not exist for road features. Internal specifications for some of these components exist within the Geography Division and within traditional mapping agencies such as the U.S. Geological Survey and the National Geospatial-Intelligence Agency. However, these specifications are often incomplete and lack the level of detail needed for precise data quality evaluations.

The situation in Europe is different. National mapping agencies such as the Ordnance Survey, Great Britain have detailed specifications for transportation networks. For examples see the OS MasterMap™ real-world object catalogue (Ordnance Survey, 2001) and the OS MasterMap Topography Layer user guide (Ordnance Survey, 2009). Efforts have been made to merge these national specifications into trans-European standards. One early attempt to create a general specification for roads data was the work done for the EuroRoadS project. This project, to build a road data infrastructure for Europe, created a specification of common road features (feature types, attribute types, and attribute value domain) and also contained specified generalization rules for the road network (Svärd, 2006). Much of this work has been superseded by the European Union's INSPIRE (Infrastructure for Spatial Information in Europe) activities. The INSPIRE Thematic Working Group on Transport Networks has produced an extensive data specification that covers major transport networks types that are defined in five distinct transport themes: Road, Rail, Water, Air transport and Cableways, including the connections between those types. The themes are defined in a way that they can be used together to support an integrated approach to transport and they may be used with other spatial data themes (INSPIRE, 2010).

Why are we concerned about detailed data specifications in a report on data quality? It is because data quality is the difference between a dataset and a "universe of discourse" (i.e. a view of the real or hypothetical world), that is defined by a product specification. Data producers and data users may use different universes of discourse and thus will have different assessments of the quality of the same dataset. A data producer's view on data quality and a user's view on data quality may merge if their requirements are identical. Product specifications that, ideally, reflect known user requirements, are key in establishing a clearly structured universe of discourse. Data product specifications are the topic of the ISO 19131 standard, *Geographic Information – Data product specifications* (ISO, 2007).

Traditionally, the Geography Division has not produced all of the content in the MTdb, but rather collected and integrated information from a variety of sources (e.g. USGS and local governments). Thus, prior to the MAF/TIGER Accuracy Improvement Program (MTAIP), there was not necessarily an urgent need to create a set of detailed content specifications and feature descriptions to accommodate the variety of information that the Geography Division acquired. Table 3 gives descriptions of the various road-type features contained in TIGER. The feature descriptions contain a certain level of ambiguity, which can cause features to be misclassified (e.g. a local road classified as a secondary road).

FEATURE NAME	MTDB COUNTS	MTFCC	FEATURE DESCRIPTION
Primary Road	20240	S1100	Primary roads are generally divided, limited-access highways within the interstate highway system or under state management, and are distinguished by the presence of interchanges. These highways are accessible by ramps and may include some toll highways.
Secondary Road	321598	S1200	Secondary roads are main arteries, usually in the U.S. Highway, State Highway or County Highway system. These roads have one or more lanes of traffic in each direction, may or may not be divided, and usually have at-grade intersections with many other roads and driveways. They often have both a local name and a route number.
Local Neighborhood Road, Rural Road, City Street	16259242	S1400	Generally a paved non-arterial street, road, or byway that usually has a single lane of traffic in each direction. Roads in this feature class may be privately or publicly maintained. Scenic park roads would be included in this feature class, as would (depending on the region of the country) some unpaved roads.
Vehicular Trail (4WD)	374252	S1500	An unpaved dirt trail where a four-wheel drive vehicle is required. These vehicular trails are found almost exclusively in very rural areas. Minor, unpaved roads usable by ordinary cars and trucks belong in the S1400 category.
Ramp	172308	S1630	A road that allows controlled access from adjacent roads onto a limited access highway, often in the form of a cloverleaf interchange. These roads are unaddressable.
Service Drive usually along a limited access highway	179600	S1640	A road, usually paralleling a limited access highway that provides access to structures along the highway. These roads can be named and may intersect with other roads.
Walkway/Pedestrian Trail	16426	S1710	A path that is used for walking, being either too narrow for or legally restricted from vehicular traffic.
Stairway	238	S1720	A pedestrian passageway from one level to another by a series of steps.
Alley	70671	S1730	A service road that does not generally have associated addressed structures and is usually unnamed. It is located at the rear of buildings and properties and is used for deliveries.
Private Road for service vehicles (logging, oil fields, ranches, etc.)	2262779	S1740	A road within private property that is privately maintained for service, extractive, or other purposes. These roads are often unnamed.
Private Driveway	414114	S1750	Private Driveway
Parking Lot Road	7643	S1780	The main travel route for vehicles through a paved parking area.
Bike Path or Trail	2301	S1820	A path that is used for manual or small, motorized bicycles, being either too narrow for or legally restricted from vehicular traffic.
Bridle Path	5	S1830	A path that is used for horses, being either too narrow for or legally restricted from vehicular traffic.

Table 3 Road/trail/path-type features in TIGER (Source: GSS Initiative Offsite, Johnson handout)

The MTAIP Technical Specifications made an attempt at creating data specifications; however, they were described in the form of contract requirements and resulted in confused or contradictory guidance. This makes the task of evaluating the data quality of this dataset more complicated.

For example, consider the instructions for collecting a "driveway" (portions of text highlighted in red for emphasis) taken from the MTAIP Technical Requirements Specification (Harris Corp., 2008, p. 115).

> **3.1.4.2.4 Driveways**
>
> A driveway is defined as an unnamed road feature of greater than 50 meters and less than 150 meters in length, which does not connect any two roads together (a dead end). Driveways may have curvature.
>
> ReqID-14567 A driveway SHALL be defined as a road or pathway that leads to a single structure where someone could live or work.
>
> ReqID-14568 Feature Improvement SHALL capture the ***Special feature class*** characteristic, to the extent described below.
>
> ReqID-8643 GIS source SHALL be used to capture ***Special feature class***, when available in the source, and according to the processing thread.
>
> ReqID-13491 Imagery source SHALL be used to capture driveways with a length greater than or equal to 50 meters and less than 150 meters, when visible in the source, and according to the processing thread.
>
> ReqID-8677 If the source file is GIS (TSC&L or commercial), Feature Improvement SHALL portray all road features captured as driveways in the source file as matched or added to the improved TIGER file.

But under the requirements for "field collected GPS data" it is noted that: "The requirements do not require field collection of driveways" (MTAIP, Vol. I, p. 81). The situation is further confused by the instructions in the MTAIP Technical Requirements Specification, Vol. II, Conflation and Portrayal Scenarios, where the instructions are:

> "Driveways are captured from imagery if they are greater than 20 meters but less than 150 meters in length. (Driveways that are greater than 150 meters in length are captured as private roads instead of driveways)." [N.B. this also is the only mention of a definition of a private road] (Source: MTAIP, Vol. II, p. 35)

Our intent here is not to be critical of the MTAIP process, but to illustrate that it is vital to have an agreed upon definition of a driveway (and every other feature) and then to adhere to that definition in the capture requirements. If adherence is not possible, then indicate in the metadata when that occurs and why. This strategy must be applied to the other components of the data specification as well. It is not possible to overemphasize the importance of specifying the "universe of discourse" in order to be able to make meaningful, quantifiable statements about data quality.

There are several types of data quality information contained in TIGER. The Spatial Metadata Identifier (SMID) identifies the source of the coordinates for road edges and provides the link

between the source and horizontal spatial accuracy information. The SMIDs were calculated using approximately 110 control points/county (located using survey grade GPS units) located at identifiable (in TIGER) road intersections. The coordinates of the location, as calculated from TIGER, were compared with the ground control, and the positional differences were summed and averaged for the geographic area in question. The error figure (e.g. 3.2 m; C95) became the SMID value for all the road edges in that portion of the dataset. Of the 46 million road edges in TIGER, approximately 38 million have a SMID value < 7.6 m (which is the desired positional accuracy for a road segment). The SMID values are reported in the FGCD compliant metadata provided with the TIGER/Line product.

The Geography Division maintains large amounts of metadata about the MTdb contents and processing history, in some cases down to the feature instance level. The internal "Business Rules Engine" performs numerous checks on attribution validity, logical consistency (including adherence to topological rules), and currency tracking. With the exception of metadata components and lineage elements, the Geography Division does not make explicit use of any of the ISO approved data quality elements or measures.

MAF Structure Points

The MAF Structure Points (MSPs) are designed to provide a latitude, longitude coordinate for a structure on the ground that corresponds to one (or more) MAF units. In addition, each operation that collects the location of a structure (such as AdCan or CAUS) can create an MSP. Thus a given structure can have multiple MSPs associated with it. Furthermore, multiple coordinates are collected for an MSP; each operation may collect two positions (i.e. two sets of latitude, longitude coordinates) for a given location. One set of coordinates reflects the GPS coordinates of the handheld device in the field. The second set are relative (manual) coordinates based on an "x marks the spot" annotation.

In terms of data quality, a series of checks are used to denote a preferred MSP. As explained by Andrea Johnson at the GSS Initiative Offsite, logical checks are performed to ensure that the MSP falls within the correct block. If two sets of coordinates exist for an MSP, the second set is designated as the alternate coordinate, with GPS coordinates taking priority over relative coordinates, however to be a "preferred" coordinate it must be located within the topology of the "MAF official" collection and tabulation blocks. If multiple GPS (or multiple relative and no GPS) coordinate MSPs exist in the official block, then the MSP from the highest ranked program/operation is preferred. It is unclear whether metadata files exist that document the lineage of each MSP.

While this series of checks places the MSP in the correct block, there are no quantitative measures (equivalent to the C95 values associated with the SMIDs) that indicate the positional accuracy of the MSPs stored in the MTdb. According to the *2010 Census Address Canvassing Lister Manual* (see Figure 2 below), the preferred location for collecting an MSP was the front door, so under ideal conditions, the MSP would be within 3m (the uncorrected accuracy of the GPS receiver) of that location. However, the field personnel did not enter a description of their location when they recorded the GPS coordinate. Post Address Canvassing clean up operations, such as the MSP Curbstone Cluster Review process, suggests that not all MSPs meet the goal of

being located near the front door. It is unclear whether spatial accuracy problems were more related to lack of adherence of procedures or lack of spatial accuracy of the features on the field personnel's map. All of these factors make it difficult to characterize the data quality of the MSPs.

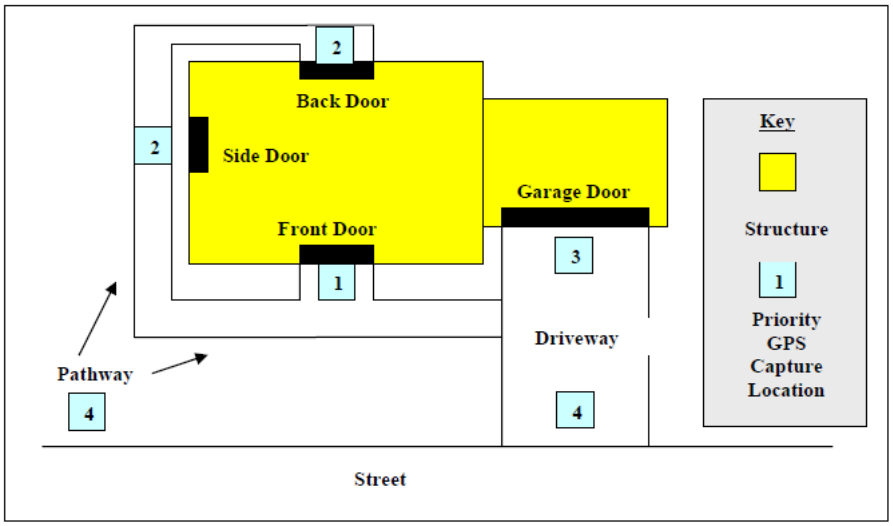

Figure 2 Locations to Collect a Map Spot (Source: U. S. Census Bureau, 2008, p. 7-13)

It is difficult to summarize this brief analysis of the current state of data specifications and quality measures. The data specifications (and resulting schemas and physical implementations) seem to have evolved over time in a bottom-up fashion. Since the Geography Division has had to accommodate data provided by numerous, often disparate, parties it is easy to see that this was an expedient, and perhaps the only viable, approach given the time and resource pressures on the program. The existing MTdb is so complex that one must ask the question whether this complexity is necessary and if it poses an impediment in implementing the goals of the GSS Initiative. In the same fashion, a comprehensive set of tests and checks have been incorporated into the processing of the MTdb content. Data sources and processes seem to be carefully tracked and recorded in a variety of metadata records. However, accessing this information is difficult and requires an intimate knowledge of the workings of the system. Some basic data quality questions are difficult to answer. How complete is the TIGER road network? How accurately are the MSPs located? Those questions can not be answered today. In the following sections, we will lay out strategies by which they can be addressed.

Evaluation of Census Requirements for Data Quality

The GSS Initiative consists of numerous operations where MTdb features or address information is replaced or updated by Geography Division staff, partners, commercial data suppliers, or Census field operations. Implicit in these operations is some notion that the new data is somehow "better" than the data it is replacing. The purpose of this phase of our research is to describe the requisite data quality characteristics needed for conducting Census operations and determine how these directly or indirectly impose requirements on the choice of data quality elements and measurements.

In testimony before the U.S. House of Representatives, U.S. Census Bureau Director Robert Groves made the following comments concerning data quality:

> *The Census Bureau strives to ensure the following tenets of quality:*
>
> - *Coverage – The MAF should ensure that each housing unit in the United States is included in the Census. Therefore, consistent and complete nationwide coverage is essential to the quality of the MAF.*
> - *Address Completeness – Each address within the MAF should have the most complete information available to ensure delivery of questionnaires and subsequent enumeration of every housing unit and group quarters within the United States. For example, in a multi-unit structure, if we do not have unique unit designations for each apartment, we will likely have problems delivering questionnaires to each address and determining who did not respond so that we can follow up to obtain an interview.*
> - *Spatial Accuracy – The accuracy of the ground location of each address in the MAF ensures that every household is counted in the correct location. The accuracy of this location also ensures the accuracy of counts, statistics, and distribution across statistical and tabulation areas.* (Groves, 2009)

Although not mentioned in his statement, one can presume that the same tenets apply to the contents of TIGER, particularly the road network and boundaries.

On the topic of data quality the draft Operational Plan for the GSS Initiative (GSS Program Element 5 – Quality Assessments) states:

> "Any number of systems and technologies can be applied to the challenges of maintaining the geographic features [and address] data in MAF/TIGER, but without a way to measure quality, and a program to help ensure quality, it can not verified that the challenges are being met effectively. This element provides the data to measure and the systems to sustain quality geographic feature [and address] data throughout the decade."

The plan also identifies the following tasks and requirements:

- Assess address quality in order to target areas for update, determine suitability for Current Surveys and other programs and refine the address universe determination for various programs.
- Identify criteria for measuring and evaluating the completeness and correctness of the address data.
- Assess the quality of the MAF/TIGER geographic features, primarily street features, in order to target areas for update and provide feedback for improving the various collection strategies employed by the Census Bureau.
- Identify criteria for measuring and evaluating the completeness and correctness of the MAF/TIGER geographic feature data.

Perhaps the most compelling statement in the plan regarding the need for data quality is:

> **"What is the impact of not including this element [quality assessments]?**
> The Census Bureau will not have suitable evidence that the MAF/TIGER [address and] geographic features data meet acceptable quality standards for program use. In addition, the Census Bureau will not be able to adequately determine whether a targeted approach to the 2020 address canvassing operation is feasible."

In distilling this stated set of high level requirements for geographic data down to its essence we conclude that the database should:

- Contain a record for every housing unit in the country,
- Contain information to ensure the delivery of questionnaires and subsequent enumeration,
- Contain a spatially accurate set of geographic coordinates to locate each housing unit,
- Contain a spatially accurate description of statistical and boundary units.

Note that these requirements relate to the ACS and decennial operations; other Census missions may have additional requirements. How are these requirements met by the data contents of the MTdb? We postulate that you need the following:

- A complete and current set of MAF Unit records to identify every housing unit
- Associated with each MAF Unit is:
    - complete and current address information[6] and
    - spatially accurate ground location (MSP)
- An address location and tabulation framework that consists of a spatially accurate and complete network of roads and boundaries.

By utilizing the basic data quality elements listed below it is possible to determine (test) if these requirements are being met:

- Completeness – account for all the feature instances
- Temporal Accuracy – feature instances are up-to-date and valid
- Positional Accuracy – spatial feature instances are accurately located
- Thematic Accuracy – feature classes, addresses, and attributes are correct
- Logical Consistency – road and boundary networks are self-consistent; the relationships between MAF Units, roads, MSPs, and boundaries are consistent among entities

In the next section, we will describe the detailed set of data quality elements and measures that can be used to evaluate addresses, roads, and MSPs.

[6] Preferably a city-style, mailable address; however other address types may be all that exist for a given MAF unit.

Develop and Document Data Quality Data Elements and Measures

In this section we describe a possible set of data quality elements and sub-elements that could be used with the MTdb. This set of data quality elements, when evaluated, will describe how well the MTdb meets the criteria described in its product specification and provide quantitative quality information. ISO TS 19138 (Data Quality Measures) is used in this effort since its objective is to guide data producers in choosing the proper data quality measures for data quality reporting, to standardize the components and structures of data quality measures, and to define commonly used data quality measures. The set of data quality elements and measures could then be utilized within the schema of the data quality report(s).

Table 4 (below) gives a set of data quality items (data quality element, data quality sub-element, data quality measure, and data quality basic measure) that have been selected for their utility with respect to the content of the MTdb. This is a subset of possible measures; over 70 measures have been documented in ISO 19138 and over 40 are in the draft FGDC address standard. Although these entries are self-explanatory, we will examine each of the 13 data quality elements with respect to how it might be applied in the case of the MTdb.

Data quality element	**Data quality sub-element**	**Data quality measure**	**Data quality basic measure**
Completeness	Commission	Number of excess items	Error count
Completeness	Omission	Number of missing items	Error count
Logical consistency	Conceptual consistency	Number of items noncompliant to the rules of the conceptual schema	Error count
Logical consistency	Domain consistency	Number of items not in conformance with their value domain	Error count
Logical consistency	Topological consistency	Number of faulty point-curve connections	Error count
Logical consistency	Topological consistency	Number of missing connections due to undershoots	Error count
Logical consistency	Topological consistency	Number of extraneous connections due to overshoots	Error count
Positional accuracy	Absolute or external accuracy	Mean value of positional uncertainties	Not applicable

Positional accuracy	Absolute or external accuracy	Mean normalized fluctuation of linear features	Database to Universe of Discourse component of the Hausdorff distance between the line in the database and the line in the universe of discourse
Temporal accuracy	Temporal validity	Number of items not in conformance with their value domain	Error count
Thematic accuracy	Classification correctness	Number of incorrectly classified features	Error count
Thematic accuracy	Classification correctness	Misclassification matrix	Not applicable
Thematic accuracy	Non-quantitative attribute correctness	Number of incorrect attribute values	Error count

Table 4 Selected data quality measures for the MTdb

Completeness

This item would be applied to virtually all of the major features and attributes in the MTdb (MAF units, addresses, road segments, MSPs, etc.). The measures describe the overcount or undercount (errors of omission or commission) of the feature in question. This measure is subject to geographic stratification; a net count of zero could mask a large overcount in one area with a large undercount in a second area.

Logical Consistency

The conceptual consistency sub-element allows evaluation and enforcement of a conceptual schema for the MTdb. In other words, integrity constraints and logical consistency can be evaluated across the entire database schema. For example, if a ZIP code change is detected by an update to the address file, then a change should be propagated automatically to the effected road edges in TIGER. Domain consistency allows for evaluation of domain values and could be used in testing the validity of address ranges associated with a road segment. Topological consistency is checked and maintained in the MTdb. However, topological consistency needs to be tested when new data is collected in the field or when data is acquired from outside sources.

Positional Accuracy

SMIDs are the MTdb's measure of positional accuracy. For the MSPs to be useful, then efforts need to be made to test their positional accuracy. The SMIDs measure the relative accuracy of a set of control points with analogous points in TIGER. The Hausdorff distance provides a measure of the fidelity of a curved line (e.g. a road) with respect to a higher accuracy representation of the same feature (see http://xw2k.nist.gov/dads//HTML/hausdorffdst.html).

Temporal Accuracy

This measure allows for the tracking of the temporal validity of a feature. For example, addresses for a sample survey should not be used unless they were validated (i.e. present in) the latest DSF updates to the MAF.

Thematic Accuracy

Classification correctness measures are closely tied to the data specification and feature instance rules. Ambiguity in the specification can result in large error reports. Are primary roads often mislabeled as secondary roads? Are the distinctions between private roads and driveways unambiguous? These measures will test for such conditions. Non-quantitative attribute correctness element measures errors in nominal attribute values (e.g. road access – limited/not limited). This is another measure that may be used in conjunction with data update and data exchange activities.

Development of Data Quality Measurement Procedures

According to methods described in ISO 19157 (Data Quality), a data quality evaluation procedure is developed through the application of one or more data quality evaluation methods. Data quality evaluation methods are be divided into two main classes, direct and indirect.

A direct evaluation method evaluates the quality of a dataset based on a comparison of the items within the dataset with reference information. The direct evaluation methods are classified as internal or external. Internal direct data quality evaluation uses only data that resides in the dataset that is being evaluated. For example, topological consistency checking of the road network is an example of an internal data quality evaluation. External direct quality evaluation requires reference data that is external to the dataset being tested. The reference data are accepted as representing the “universe of discourse” or, in other words, “ground truth”. In this case, an example would be that the positional accuracy of TIGER roads might be evaluated by comparing the roads data with the roads visible in high resolution orthoimagery, or field personnel with high precision GPS devices could retrace the road segment. For both external and internal evaluation methods, one of the following inspection methods may be used:

- full inspection or
- sampling.

A “full inspection” tests every item in the population and is most appropriate for small populations or for tests that can be accomplished by automated means. Sampling tests are performed on subsets of the geographic data. An informative annex in 19157 provides information on how to select an appropriate sampling strategy.

An indirect evaluation method infers or estimates data quality using information about the data such as lineage information. This type of external knowledge may include, but is not limited to, data quality overview statements or other quality reports on the dataset or data used to produce

the dataset. It may be estimated from knowledge about the source, tools, and methods used for capturing the data and evaluated against procedures and specifications that represent the "best practice" for this product. If indirectly evaluated data quality has been reported, it should be accompanied by a description on how it was determined. In some cases it might be misleading or not even possible to report indirectly evaluated data quality as quantitative results. In those cases the data quality may be described in textual form using a descriptive result. As such, it is often subjective. For these reasons direct evaluation methods are preferred to indirect evaluations.

Table 5, below, provides some examples of the use of the direct evaluation method described in ISO 19157.

DQ_ element	DQ_ subelement	DQ_ measure	DQ_ conformcelevel	DQ_ EvalProcedureDesc	DQ_ value
Completeness	Omission	Number of missing items	Declared Quality Level (DQL) 2,5% (based on ISO 2959-4 standard)	Stratified random sampling using inspection of items in the field. Sample size 32 items	Accepted (Number of missing items 1)
Logical consistency	Domain consistency	Number of items not in conformance with their value domain	0	Full inspection	Rejected (Number of wrong items 1)
Positional Accuracy	Absolute of external accuracy	RMSE	2 m	Random sample	Accepted (RMSE 1,5 m)

Table 5 Examples of the usage of ISO 19157 Standard (Source: Jakobsson and Giversen, 2007, p.22)

Data quality evaluation procedures can be applied at various points in a product's life cycle. At the production stage, the producer may apply quality evaluation procedures as part of quality control during stages of the production process. The description of the applied quality evaluation procedures, when used for production quality control, may be reported as lineage metadata including, but not necessarily limited to, the quality evaluation procedures applied, conformance quality levels established, and the results.

A quality evaluation process might be used upon completion of production to produce and report data quality results. These results may be used to determine whether a dataset conforms to its product specification or not. If the dataset passes inspection (composed of a set of quality evaluation procedures), the dataset is considered to be ready for use.

Most data producers, including the Geography Division, apply quality evaluation procedures as part of data processing procedures. However, formal documentation of these quality processes is not usually performed. Among the national mapping agencies, the Institute Geographic National, France has been one of the most thorough in creating precise product specifications, documenting quality processing, and performing field based quality testing.

Note that the data quality measurements provide the data for the user to make an informed decision about the disposition of the data. Data sets or data elements may be rejected or accepted based on defined evaluation criteria or thresholds. It will be incumbent on the Census Bureau to define the appropriate acceptance criteria.

The results of data quality evaluation procedures may be reported using metadata or in a standalone report. ISO 19115 and 19157 contain protocols (as well as UML models and data dictionaries) for reporting this information. Additionally, a standalone report may be created to provide more details than are required in the metadata. There is no specified format for this report. However in accordance with the standards, the standalone report cannot replace the metadata. The metadata should provide a reference to the standalone report when it exists.

ISO/TC 211 has another standard under development that may provide additional guidance in this area. ISO 19158 *Geographic Information – Quality Assurance of Data Supply* is intended to provide a framework for quality assurance specific to geographic information. It is based upon the quality principles and quality evaluation procedures identified in ISO 19157 and upon the general quality management principles as defined in ISO 9000:2005. This document is scheduled to be completed and published in 2011.

A set of data quality measurement procedures were developed for use in the 2010 Address Canvassing Operation. Census Memorandum Series #F-01R1, *2010 Census: Revised Quality Control Plan for the Address Canvassing Operation,* lays out a comprehensive, direct, external evaluation of the street and address information collected during Address Canvassing.

For the map features (streets), the memo states:

> "To review the quality of the map updates, we will conduct a quality evaluation on map features (streets) added or deleted by the Listers. We will not verify spatial accuracy. We will only verify that the updated feature exists or does not exist.
>
> A separate QC staff will conduct a quality control field check on every AA. During this field check, the QC Lister will check a single map feature update. The feature update QC will take place at the same time as other QC checks that are part of the DQC.
>
> To accomplish this, the HHC will first check whether there are any added or deleted map features in the AA. If there are, the HHC will randomly select one of those features for verification by the QC Lister, then the HHC will select a starting point for DQC in the same block as the selected feature. If there are no added or deleted map features in the AA, the HHC will skip the MFQC entirely and will select the DQC random start point from a randomly selected block in the AA."

For addresses a 3% sample of the address records were tested as follows:

> "At the completion of production listing for every AA, a QC Lister will check a random sample of consecutive address records in the AA. It is important to understand that the QC sample is representative of the entire AA. If there are errors in the QC sample then we assume there are about the same rate of errors in the remainder of the AA. Therefore, if the QC sample fails, the entire AA fails. For any AA that fails the DQC, the remainder of the address records in that AA will be checked.
>
> To carry out the DQC, the QC Lister will review listings that have been verified, listings that have been added, and listings that contain corrections. The DQC is designed to identify significant errors and provide an assessment of each work assignment. The DQC is done concurrently with the DV."

A review of the location of the MSPs was also part of the original process. The procedures to be used were as follows:

> "…the QC Lister…will review the map spot associated with the address record he or she is working on in the Block Map.
>
> 1. The QC Lister will look at the associated map spot (labeled with its map spot number) and determine if it is in the correct place on the map.
> 2. If the map spot information is correct, the QC Lister will select **Accept** and QC is finished for this address record.
> 3. If the map spot information is not correct, the QC Lister will
> a. Select the **Reject** button, and
> b. Re-do the map spot placement using standard map spot placement procedures."

However, this procedure was removed from the process due to unexpected impacts in the canvassing schedule caused by the repositioning of the MSPs.

The information resulting from the quality control operation was not shared with the consultants. It is possible that an analysis of the results could be very useful to the Geography Division in establishing a set of quality baseline statistics for MAF/TIGER.

Identification of Relevant Data Quality Components for Data Exchange

The inclusion of data quality information with address and/or spatial data sets is critical when data is exchanged between partners. Knowing the quality of a dataset is a key factor in MTdb update operations. Data in the MTdb should only be replaced if a "better" component is used for the update. Data quality measures are critical in determining if data from outside sources (or data generated by internal operations) are appropriate for revision of the MTdb. In evaluating data quality components for data exchange, the process is identical to that used in evaluating one's own holdings. This process is diagramed in Figure 3 (below).

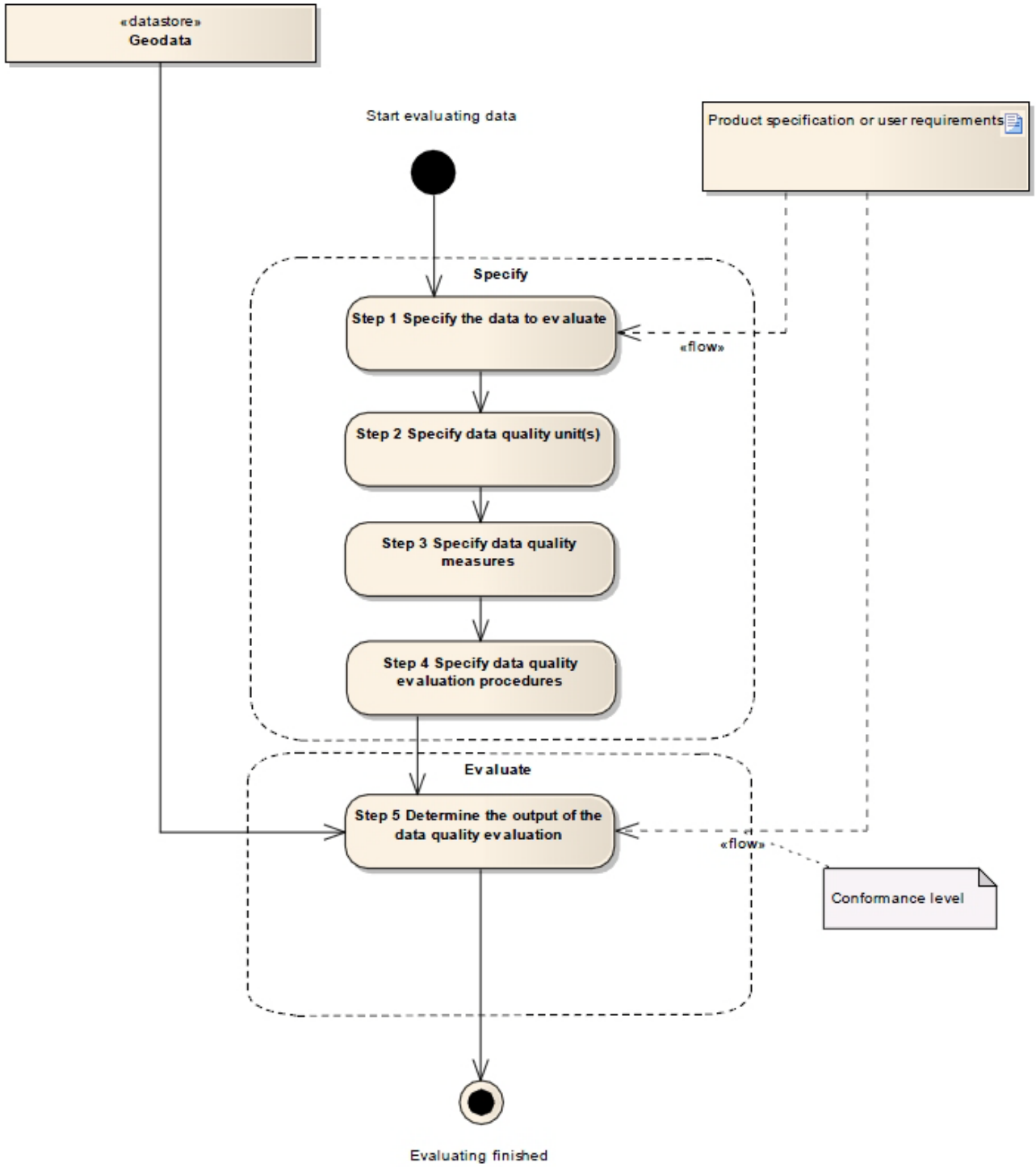

Figure 3 Process steps for evaluating data quality (Source: ISO, 2010, p. 18)

If data partners have quality information of the type being described in this report, then evaluation of datasets from that partner could be relatively straightforward. However, given the resources needed to create complete data quality reports, it is possible (or more likely probable) that data partners will be unable to provide data quality elements and measures that we have specified in this report. Then Census will have to employ quality evaluation procedures to establish if a dataset meets the conformance quality levels specified in Census's data requirements. Indirect as well as direct methods may be used in analyses of dataset conformance to Census requirements.

Each of the data quality elements listed previously in Table 4 need to be considered when evaluating a contributed data product. If a data quality element fails to meet the desired conformance level this does not necessarily cause the dataset to be rejected. Census will need to judge whether acceptable (or superior) conformance in some elements (e.g. completeness or temporal validity) outweigh lack of conformance in another element (e.g. positional accuracy). For example, a contributed dataset might contain the street names, addresses, and street geometry for a new subdivision which is missing in the MTdb. However the positional accuracy of the streets does not meet the 7.6m criteria of the Geography Division. In such a case then a quality assurance unit (assuming one existed within the Geography Division with the objective of performing data evaluation and correction) could measure and test the data and bring it into compliance or return the deficient datasets to the data partner for correction.

It is important to note that data specifications are a key component of the data exchange process. MTdb data that are in compliance with the data specifications of Geography Division must be brought into concordance with that of the data exchange partner. Consider the image of a portion of Charleston County, South Carolina in Figure 4 below. The red lines represent roads (all types) from TIGER Line 2009. The yellow lines represent roads from the Charleston E911 authority. Without detailed data specifications and data attribution, reconciling these two datasets will be very difficult. These topics will be discussed in our report *Researching Address and Spatial Data Digital Exchange and Data Integration.*

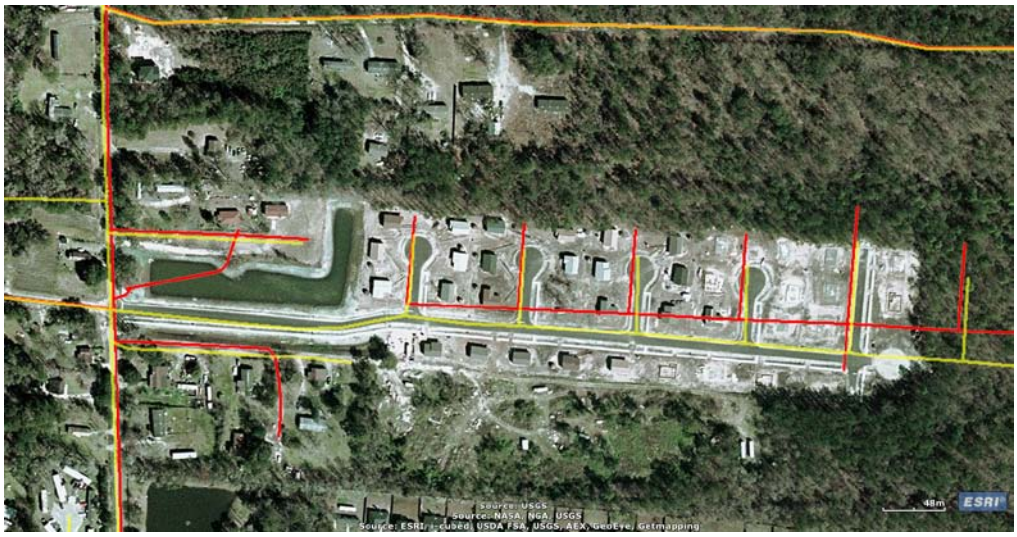

Figure 4 TIGER Line 2009 and Charleston County E911 Roads

Relating Data Quality Procedures to Applicable Support and Update Operations

Various components and operations of the Census Bureau (such as Address Canvassing) provide data to the Geography Division for inclusion into the MTdb. In the past, these data sets have been accompanied with a process history, but generally these data are provided with little or no specific data quality information. As part of future data collection operations, data quality information should be collected as well.

Our analysis in *Reporting on the Use of Handheld Computers and the Display/Capture of Geospatial Data,* determined that field based data collection to support the GSS Initiative falls into the general categories of collecting attribute information (e.g. addresses and street names) and locational data (e.g. housing unit locations and coordinates of streets). In the Address Canvassing operations for the 2010 decennial census, the field personnel were instructed to:

- "Compare what is on the ground to what is in the Address List.
    - Address Canvassing is conducted by looking at what is on the ground and comparing it to what is in your Address List. We call this 'ground-to-list.'
    - If there are new or existing homes on the ground, but not in the Address List, add them.
    - If there are addresses on the Address List, but not on the ground, delete them.
    - The Address List should reflect exactly what is on the ground.

- Update the census maps on the hand held computer.
    - Make corrections to street names that currently exist on the map.
    - Delete streets that do not exist.
    - Add streets you find on the ground that are currently not on the maps on the HHC.

- Update the status of every address in each block of your AA. Address status choices include:
    - Housing Unit - includes under construction and vacant trailer sites.
    - Other Living Quarters
    - Does Not Exist
    - Duplicate
    - Nonresidential
    - Uninhabitable

- Update the Address List
    - Each living quarter in the block you are canvassing should be represented on the Address List in the HHC. If you do not find the address on the list, add it.
    - Confirm the address either by observation or a brief interview at the living quarter using the guidelines found in Chapter 2, topic 5 of this manual.
    - Update the address as prompted by the HHC.
    - Give a D-31, Confidentiality Notice to each person you speak to when gathering information to update the Address List.
    - After completing one block, continue in another block until the AA is finished.

- Collect a map spot at each valid living quarters.
  - Use your HHC and GPS to collect a map spot when you visit a living quarters in your assignment.
  - EXCEPTION: For multi-unit buildings such as apartments, only collect a map spot for the first unit in the structure."

(Source: Census Bureau, 2008, pp 1-12, 13)

These operations can be generalized into verification and update activities applied to four generic classes of data:

- Housing unit attributes,
- Housing unit locations,
- Street attributes, and
- Street locations.

The relationships between these field data classes and the applicable data quality elements and sub-elements (which are the same elements as previously listed in Table 4) are described in a matrix form in Table 6. If the data quality elements (and the associated sub-elements, quality measures, etc.) are relevant to the field data class, the cell of the matrix is marked with an "x", if not it is marked "n/a".

Data quality element	Data quality sub-element	Housing Unit Attributes	Housing Unit Location	Street Attributes	Street Location
Completeness	Commission	x	x	x	x
Completeness	Omission	x	x	x	x
Logical consistency	Conceptual consistency	x	n/a	x	n/a
Logical consistency	Domain consistency	x	x	x	x
Logical consistency	Topological consistency	n/a	x	x	x
Logical consistency	Topological consistency (undershoots)	n/a	n/a	n/a	x
Logical consistency	Topological consistency (overshoots)	n/a	n/a	n/a	x
Positional accuracy	Absolute or external accuracy	n/a	x	n/a	x
Positional accuracy	Absolute or external accuracy (linear feature)	n/a	n/a	n/a	x
Temporal accuracy	Temporal validity	x	x	x	x
Thematic accuracy	Classification correctness	x	n/a	x	n/a

Thematic accuracy	Classification correctness (matrix)	x	n/a	x	n/a
Thematic accuracy	Non-quantitative attribute correctness	x	n/a	x	n/a

Table 6 Relationship between data quality elements and field data classes

This matrix indicates the set of data quality information that a field operation should provide to the Geography Division for each class of data. The set of required data quality information would be the same if non-field operations (e.g. photo interpretation of imagery) were used as part of the data acquisition and/or update process.

The *2010 Census: Revised Quality Control Plan for the Address Canvassing Operation* included processes to collect most of the information types noted in this matrix. The major missing components were measures of positional accuracy with respect to the MSPs and street locations. No information was available to the consultants that indicated how the 2010 quality control information might be conveyed to the Geography Division, or how it would be used upon receipt. Nonetheless, the AdCan quality control plan (augmented with spatial accuracy components) provides a sound basis for creating a quality control plan for use in forthcoming field operations in CAUS and perhaps ACS.

Summary and Recommendations

The Census Bureau has as one of its tenets the desire to have the MTdb be complete, accurate, and spatially precise. The success of the GSS Initiative hinges on the ability to demonstrate that the MTdb contents are of a high enough quality that targeted address canvassing can be used in the 2020 decennial census. As was noted in the GSS Initiative draft operational plan, without a way to measure quality, and a program to help ensure quality, it cannot be verified that the challenges to maintain the geographic features and address data in the MTdb are being met effectively.

As noted by the National Research Council:

> "Having made serious investments in upgrading the technical platform of the Bureau's geographic systems during the 2000–2010 decade, the challenge now becomes one of keeping those resources up to date in the most accurate way possible. An original focus of the MAF/TIGER Enhancements Program of the previous decade was on quality metrics (methods for assessing the quality and geographic accuracy of both MAF and the line features in the TIGER database) and update mechanisms…Generally, the Census Bureau would benefit from a program of field spot-checks, comparison with third-party sources (including addresses drawn from administrative records data files), and the like in order to have continuous diagnostic measures of the quality of MAF and TIGER and to detect priorities for update and maintenance." (Source: NRC, 2010, p.82)

It is evident that it is incumbent upon the Geography Division to have in place the information and processes that can be used to perform the necessary data quality assessments.

To achieve this goal, we offer the following recommendations:

- Adopt and utilize the language, structures, and information infrastructure for data quality evaluations as promulgated by the ISO 19100 suite of standards and by the FGDC.

- Review and revise as necessary the data content specifications for the MTdb so that the elements are precisely and formally defined in terms of the requirements that they are intended to fulfill.

- Validate and revise as necessary the proposed set of data quality elements against Census user requirements.

- Determine appropriate conformance levels for each data quality measure.

- Choose what data quality evaluation procedures will be employed and when in the data product life cycle they will be applied.

- Establish an initial set of data quality values for the MTdb by:

  - Processing and analyzing the existing metadata holdings and placing the appropriate contents into ISO 19100 compliant form and

  - Obtaining and analyzing the results from the Quality Control Plan for the Address Canvassing Operation to create a baseline set of data quality measures.

- Utilize data specifications, data quality measures, and conformance levels in evaluating potential data exchanges/acquisitions with third parties.

- Utilize data evaluation software as part of the data production process, particularly in handling contributions from data exchange partners.

- Apply data quality procedures and processes to field update operations.

- Other Census groups use MTdb data for various purposes and these groups have their own quality assurance procedures. When MTdb data fail the users, there needs to be a feedback loop from the users to the Geography Division so that GEO can learn how to correct the errors. In essence, GEO should be made aware of what data (elements) failed and test them against their metrics to see where the metrics failed them and what corrective action is needed.

- Commercial software exists to perform testing and reporting in conformance with ISO 19100 standards. Examples of such software include ArcGIS DataReviewer and 1Spatial Radius Studio. The Geography Division should evaluate this type of software for its applicability to Census operations.

Over the last two decades the Geography Division has, through the creation of the MTdb, assumed the leadership role within the Federal government as the source of authoritative, complete, multipurpose geospatial data for the Nation. To maintain that leadership role, the Geography Division will need to continually enhance the content of the MTdb, and more importantly, provide quantitative evidence to its users (both internal and external to the Bureau) that the data are of high quality and fit for multiple uses. This is a challenging task, but one that, based on its past record, the Geography Division can meet. Adopting a systematic approach to implementing a data quality management program is critical to achieving this goal.

References

Federal Geographic Data Committee (FGDC). 1998. *Content Standard for Digital Geospatial Metadata* (CSDGM), Ver. 2 (FGDC-STD-001-1998). Reston, VA: FGDC. Available online at http://www.fgdc.gov.

Federal Geographic Data Committee (FGDC). 2010. *United States Thoroughfare, Landmark, and Postal Address Data Standard (Draft).* Reston, VA: FGDC. Available online at http://www.fgdc.gov.

Goodchild, M., and S. Gopal. 1989. *Accuracy of Spatial Databases.* Bristol, PA: Taylor and Francis Inc.

Groves, R. 2009. *2010 Census: Master Address File, Issues and Concerns.* Prepared statement before the Subcommittee on Information Policy, Census, and National Archives, U.S. House of Representatives, October 21, 2009.

Guptill, S.C., and J.L. Morrison. 1995. *Elements of Data Quality*. New York, NY: Elsevier Science Ltd.

Harris Corporation. 2008. *Technical Requirements Specification for MAF/TIGER Accuracy Improvement Project (MTAIP)*, 2 Volumes. Prepared for U.S. Census Bureau Contact NO: 50-YABC-2-66005. April 21, 2008.

INSPIRE (Infrastructure for Spatial Information in Europe). 2009. *INSPIRE Data Specification on Addresses – Guidelines.* Brussels: INSPIRE. Available online at: http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_AD_v3.0.pdf

INSPIRE. 2010. *INSPIRE Data Specification on Transport Networks – Guidelines.* Brussels: INSPIRE. Available online at: http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_TN_v3.1.pdf.

ISO (International Organization of Standardization), Technical Committee 211. 2002. *ISO 19113, Geographic Information – Quality principles*. Geneva, Switzerland: ISO. Available online from http://www.isotc211.org/.

ISO. 2003a. *ISO 19114, Geographic Information – Quality evaluation procedures*. Geneva, Switzerland: ISO. Available online from http://www.isotc211.org/.

ISO. 2003b. *ISO 19115, Geographic Information – Metadata.* Geneva, Switzerland: ISO. Available online from http://www.isotc211.org/.

ISO. 2006. *ISO 19138, Geographic Information – Data quality measures.* Geneva, Switzerland: ISO. Available online from http://www.isotc211.org/.

ISO. 2007. *ISO 19131, Geographic Information – Data product specifications*. Geneva, Switzerland: ISO. Available online from http://www.isotc211.org/.

ISO. 2009. *ISO 19114, Geographic Information – Quality evaluation procedures*. Geneva, Switzerland: ISO. Available online from http://www.isotc211.org/.

ISO. 2010. *ISO 19157, Geographic Information – Data Quality*. Geneva, Switzerland: ISO. Available online from http://www.isotc211.org/.

ISO/TC211 Geographic Information/Geomatics. 2009. *Standards Guide*. Geneva, Switzerland: ISO. Available online at http://www.isotc211.org/.

Jakobsson, A., and J. Giversen. 2007. *Guidelines for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies*. Brussels: Eurogeographics. Available online at: http://www.eurogeographics.org.

Jensoulin, R., and M. Goodchild. 1998. *Data Quality in Geographic Information, From Error to Uncertainty.* Paris: Hermes.

NRC (National Research Council). 2010. *Envisioning the 2020 Census.* Washington, D.C.: National Academy Press.

Ordnance Survey. 2001. *OS MasterMap™ real-world object catalogue.* Available online at: http://www.ordnancesurvey.co.uk.

Ordnance Survey. 2009. *OS MasterMap™ Topography Layer User Guide.* Available online at: http://www.ordnancesurvey.co.uk.

Svärd, T. 2006. *Final specification of core European road data*. Stockholm: National Land Survey of Sweden. Available online at: http://www.euroroads.org/php/documents.php

U.S. Census Bureau. 2008. *Address Canvassing, Lister Manual.* Suitland, MD: U.S. Census Bureau.

Wenzhong, S., P. Fisher, and M. Goodchild. 2002. *Spatial Data Quality.* New York, NY: Taylor and Francis Inc.