

BEYOND POSITIONAL ACCURACY IMPROVING GEOSPATIAL METADATA STANDARDS

Frederick R. Broome and Leslie Godwin
U.S. Bureau of the Census

ABSTRACT

Executive Order 12906 instructs each Federal agency to document using the Federal Geographic Data Committee's (FGDC) geospatial metadata standard* all new and, to the extent applicable, existing geospatial data it collects or produces. The Census Bureau identifies both its geographic feature sets such as TIGER/Line and its georeferenced statistical data sets as geospatial data, making the geospatial metadata standard the applicable data documentation metadata standard for the majority of the Bureau's data sets. The geospatial metadata standard allows for the complete documenting of geographic feature sets. The Bureau's many georeferenced statistical data sets, however, often contain indirect spatial referencing. The emphasis of these data sets is thematic, or topical, in nature. The ability to completely document the thematic aspect of these data sets is of importance to the Bureau and to data users.

The Department of Commerce is the FGDC member agency responsible for developing cultural and demographic data set standards. The Census Bureau represents the Department in this area by chairing the FGDC's Subcommittee on Cultural and Demographic Data. The subcommittee interprets their standards responsibilities as extending to metadata standards and has identified four areas of particular importance to cultural and demographic data sets they feel are not wholly addressed by the geospatial metadata standard. These areas are: 1) themes, 2) the geographic dimension of the data, specifically indirect geospatial (nominal) coding, 3) data quality, and 4) the temporal dimension. The subcommittee has proposed a "Content Standards for Cultural and Demographic Data Metadata" allowing for the creation of metadata to identify the cultural and demographic component of data sets and facilitating the production of geospatial metadata. This paper introduces the proposed cultural and demographic data metadata standard and explores its relationship to the existing geospatial metadata standard.

KEYWORDS

Cultural and Demographic, Federal Geographic Data Committee, Geospatial Data, Metadata, Standards

1. INTRODUCTION

Executive Order 12906, signed in April, 1994, instructs each Federal agency to document all new and (to the extent applicable) existing geospatial data it collects or produces using the standards under development by the Federal Geographic Data Committee (FGDC) and to make the standardized documentation (metadata) electronically accessible. The FGDC's "Content Standard for Digital Geospatial Metadata," issued in June, 1994, is the data documentation standard referenced in the executive order. The Census Bureau identifies both its geographic feature sets such as TIGER/Line and its geo-referenced statistical data sets as geospatial data. As a result, the FGDC's geospatial metadata standard is the applicable data documentation standard for metadata for the majority of the Bureau's data sets.

The geospatial metadata standard covers many aspects of data sets and allows the complete documenting of geographic feature sets. Many of the Bureau's data sets, however, are statistical data sets containing indirect spatial referencing. These data sets are thematic, or topical, in nature. Data users, including decision makers and policy makers at all levels of government, most often utilize the topical aspect of data. Being able to completely document this aspect of its data is of importance to the Bureau.

The Department of Commerce is the FGDC member agency responsible for developing cultural and demographic data set standards. This responsibility was passed on to the Census Bureau (an active participant of the FGDC) and includes chairing the FGDC's Subcommittee on Cultural and Demographic Data. The subcommittee interprets their standards responsibilities as extending to metadata standards. The subcommittee has identified four areas of particular importance to cultural and demographic data sets they feel are not wholly addressed by the geospatial metadata standard. These areas are: 1) themes, 2) the geographic dimension of the data, specifically indirect geospatial (nominal) coding, 3) data quality, and 4) the temporal dimension. A further area of importance to the Bureau, not initially identified by the subcommittee, is the statistical aspect of the data.

The subcommittee has proposed a "Content Standards for Cultural and Demographic Data Metadata" allowing for the creation of metadata to identify the cultural and demographic component of data sets and facilitating the production of geospatial metadata. This paper introduces the proposed cultural and demographic data metadata standard and explores its relationship to the existing geospatial metadata standard. Benefits of the acceptance of this

standard for completely documenting Census Bureau data sets are identified.

2. THE NATURE OF CULTURAL AND DEMOGRAPHIC DATA

2.1 The Diversity of Cultural and Demographic Data

Many acknowledge the data they both produce and use has a "geospatial" aspect. by the Federal Geographic Data Committee (FGDC)¹ and, more recently, by Executive Order 129062 is:

information that identifies the geographic location and characteristics of natural or constructed features and boundaries of the earth. This information may be derived from, among other things, remote sensing, mapping, and surveying technologies. Statistical data may be included in this definition at the discretion of the collecting agency.

The FGDC recognizes and coordinates categories of geospatial data. The FGDC is comprised of eleven category-based subcommittees, including the Subcommittee on Cultural and Demographic Data (SCDD)³. FGDC subcommittees coordinate activities relevant to their geospatial data categories.

When asked to categorize their data, "cultural and demographic" is not the primary selection of most data producers or consumers. The type of data belonging in this category remains vague. Cultural, the more general of the two terms, appears so general as to be unbounded; while, in contrast, demographic appears so precise as to be limiting. What types of data are cultural and demographic? The SCDD charter defines it as follows:

Cultural and demographic geospatially referenced data include, but are not limited to, compilations of information about the people and institutions of the United States and its territories: the characteristics of the people, the nature of the dwellings in which they live, the economic activities they pursue (such as the farms on and establishments or organization for which they work), the facilities they use to support their health and recreational needs, the environmental consequences of their presence, and the boundaries, names and numeric codes of the geographic entities used to report the information collected.

Cultural and demographic data is a fairly broad category of data which centers on the "human dimension" and consists of any data related to humans, their activities, and their interaction with the environment. Much data which is specifically addressed by other

FGDC thematic subcommittees can in a general sense be considered cultural and demographic, but are better coded by a standard specific to that theme. However, the important issue is that the majority of these data have common characteristics essential to the successful sharing and exchange of the cultural and demographic component of the data. It is through the sharing that efficiencies in data collection operations occur and new insights in data relationships become possible.

2.2 A Common Characteristic

Geospatial data have many characteristics by which they may be viewed. A characteristic common to and shared by all geospatial data sets is a spatial or positional component. Other characteristics of data important to data producers and consumers include (1) a data set's source or lineage, (2) the quality of the data set, (3) a data set's temporal aspect, and (4) a data set's topical aspect.

The SCDD believes the topical aspect of geospatial data sets to be the characteristic defining cultural and demographic data sets. More importantly, the SCDD feels it is the topical or thematic component of cultural and demographic data which makes these data sets valuable to policy makers, decisions makers, and similar metadata users.

3. METADATA STANDARD FOR CULTURAL AND DEMOGRAPHIC DATA

3.1 The FGDC's Existing Metadata Standard

The FGDC approved "Content Standards for Digital Geospatial Metadata" (a data Standard) standardizes information that helps prospective data set users determine what data exists, the fitness of these data for applications, and conditions for accessing the data by specifying the information in the metadata⁴ for a set of digital geospatial data. The Geospatial Metadata Standard provides a common set of terminology and definitions for concepts related to geospatial metadata. Executive Order 12906 instructs Federal agencies that by 1995 they are to use the Geospatial Metadata Standard to document new geospatial metadata and to provide these metadata to the public through the National Geospatial Data Clearinghouse to further the goals of the NSDI and the National Information Infrastructure (NII).

Each FGDC subcommittee is responsible for developing data content, metadata, and data exchange standards for their particular category of geospatial data. The SCDD has reviewed the Geospatial Metadata Standard's ability specifically to describe cultural and demographic data sets and finds that numerous aspects of cultural and

demographic data can be thoroughly described. The Geospatial Metadata Standard provides an excellent means of documenting the positional component of data sets such as political or statistical area boundaries.

Unfortunately, decisions based on geospatial thematic data rarely require information having high positional accuracy (though it is important to remember the need for positional accuracy increases as activities move from the policy level to the implementation level.) Few data users and decision makers consider cultural and demographic data primarily in terms of its positional characteristics; most decision makers analyze the topical aspects of the data in a topological sense rather than a topographical one; e.g. "near" or "next to" rather than "300 meters from". The means of describing the component of data most common to cultural and demographic and most important to many data users is not found in metadata created using the Geospatial Metadata Standard. It is difficult for metadata producers to use the Geospatial Metadata Standard to adequately describe the topical aspect of their cultural and demographic data.

3.2 A SCDD Proposed Metadata Standard

The SCDD has proposed a "Content Standard for Cultural and Demographic Data Metadata" (CDD Metadata Standard) to provide a more thorough and useful documentation of cultural and demographic data achievable than by using the Geospatial Metadata Standard alone; and to do so while not losing the spatial nature of the data. The proposed standard is designed to augment, not replace, the existing standard. The proposed CDD Metadata Standard allows data producers to present the thematic perspective of their data while retaining its positional nature. Indeed, the two standards are harmonized to the point of sharing over 80% of the primary element terminology, thus providing a direct crosswalk between the two for easy search.

4. THE PROPOSED CDD METADATA STANDARD

4.1 Similarities

The CDD Metadata Standard was developed with two objectives: (1) to allow for the creation of metadata to better identify the thematic component of data sets, and (2) to facilitate the production of geospatial metadata. To accomplish the second of these goals, the Geospatial Metadata Standard was used as a foundation for building the CDD Metadata Standard. The proposed standard has a look and feel familiar to data producers and consumers already using the

Geospatial Metadata Standard. Similarities between the standards include shared terminology, metadata compound and data elements, production rules, and document organization.

The SCDD Metadata Standard contains four sections or parts. The first two parts are informational, explaining the developmental background (Part 1: Introduction to the Standard) and presenting the intent of the proposed standard (Part 2: Modeling the Data and the Standard).

Part 3 of the proposed standard (Part 3: Content Standard) is a "working" section essential for producing cultural and demographic data metadata. Part 3 is organized into metadata components including a hierarchy of compound elements and data elements that are the cultural and demographic data metadata content. As defined in the Geospatial Metadata Standard and applied in both standards, a compound element is a group of data elements and other compound elements. All compound elements represent higher-level concepts that cannot be represented by individual data element; they are described, either directly or indirectly, by data elements. A data element is a logically primitive item of data. For its compound and data elements, each standard specifies a number or identifier, definition, type, domain, constraint, and cardinality. The CDD Metadata standard assigns the data elements it shares with the Geospatial Metadata Standard unique numbers or identifiers but maintains the remainder of the characteristics as assigned in the Geospatial Metadata Standard (see Figure 1). This was done for structured purposes and technologically poses no burden on the user.

Shared production rules structure geospatial and cultural and demographic metadata in a similar manner. These rules describe for compound elements and data elements (1) "replacing", "producing", and "consisting of principles", and (2) "selection" (or "exclusion") principles. The characteristics "constraint" and "cardinality" may alternately be considered as the production rules (3) "iterations", and (4) "optionality".

4.2 and Differences

4.2.1 A New Production Rule to Extend the Descriptive Potential of Metadata

The SCDD determined that an additional production rule will greatly aid metadata producers in describing the complexity of their cultural and demographic data sets. The CDD Metadata Standard allows metadata producers to identify multiple entries of data elements as either same level elements or, alternately, hierarchically nested

elements. Many cultural and demographic data sets contain both distinct and layered levels of data. These differing data may be of greater or lesser importance to the data producer and have received a differing emphasis. The ability to pass this information on to the metadata user is possible with the adoption of hierarchical ordering for a limited number of data elements.

GRAPHIC

Figure 1. Comparing characteristics of a shared data element.

4.2.2 A New Data Element To Extend Versatility

A new, "free-floating" data element the SCDD has named as "tag" in the CDD Metadata Standard allows the metadata producer to include additional information or identifiers for internal reference purposes at any location in the metadata. Tag information need not be explained to nor necessarily understood by the general consumer of the metadata. Tag allows the metadata producer to interpret or translate the common set of metadata terminology and definitions to the metadata producer's unique terminology within the metadata itself. Metadata users may ignore Tag data elements and successfully obtain information necessary to evaluate the fitness of use of a cultural and demographic data set.

4.2.3 An Emphasis on the Topical Aspect of Data

The proposed CDD Metadata standard defines six mandatory components of cultural and demographic data metadata: (1) Identification/General Information, (2) Themes, (3) Geographic Information, (4) Temporal Information, (5) Source/Lineage Information, and (6) Data Quality Information. Four of these components--themes, geographic, data quality, and temporal--are central to the topical nature of cultural and demographic data, with themes providing integral information. Figure 2 depicts the metadata components of the proposed and existing standards.

GRAPHIC

Figure 2. The components of the metadata standards.

The primary focus of and differences between the standards are more clearly seen by slightly rearranging and annotating Figure 2 (see Figure 3). The SCDD recognizes the geospatial components of the Geospatial Metadata Standard to be so comprehensive that they are neither augmented nor addressed in detail in the CDD Metadata Standard. Entity and Attribute Information, Spatial Reference Information, and Spatial Data Organization Information components

are unique to the Geospatial Metadata Standard. The CDD Metadata Standard provides a more thorough picture of the thematic aspects of cultural and demographic data sets; this is accomplished by expanding upon concepts found primarily in the Identification and Data Quality components of the Geospatial Metadata Standard. Figure 3 compares the components of the proposed and existing standards.

5. CREATING METADATA FOR THEMATIC DATA SETS: AN EXAMPLE

The differences in documenting the thematic aspect of data can best be illustrated by providing an example of metadata for the thematic component of a cultural and demographic data set using both the existing and proposed standards.

The Bureau of the Census conducts a Census of Agriculture every fifth year, collecting data in years ending in either two or seven. One table or, for the purposes of this article, data set, presents data on the number, land, and value of farms at two year intervals for states by county (see Figure 4).

Policy and decision makers and other data consumers may be interested in differing aspects of this data depending on their current or prospective needs. A realtor may wish to know about the appreciating or depreciating value of farm land in a community; a planner may need information on the changing percentage of land area in farms for developing land use and zoning policies; an economist may be interested in the number of farms with sales of \$2,500 or more; a prospective or retiring farmer may be researching the average value per acre of farm land to plan his/her future.

GRAPHIC

Figure 3. Comparing the components of the metadata standards.

GRAPHIC

Figure 4. Overview of a thematic data set.

The Geospatial Metadata Standard provides the metadata producer with the compound element "Theme" in its Identification component. The Geospatial Metadata Standard acknowledges the complexity of data sets by allowing the metadata producer to provide multiple instances of Theme. Theme consists of two data elements. "Theme Keyword" is a common-use word or phrase used to describe the subject of the data set, this data element may also be repeated as many times as desired. The domain of Theme Keyword is free text;

the standard suggests subject/index term sources for selecting Theme Keyword(s). An associated data element is "Theme Keyword Thesaurus" which allows the metadata producer to reference a formally registered thesaurus or a similar authoritative source of the identified theme keywords.

The CDD Metadata Standard raises the visibility of the thematic aspect of data sets by providing a Theme component. It acknowledges that many data sets are complex and consist of multiple main themes: the compound element "Main Theme" is the equivalent of the Geospatial Metadata Standard's "Theme" compound element and the metadata producer may provide multiple instances of Main Theme.

Main Theme consists of five data elements.

"Definitions/Classifications" is similar to and expands the concept of Theme Keyword Thesaurus, allowing the metadata producer to either reference a formally registered thesaurus or a similar authoritative source for the identified theme keywords or provide verbal descriptions defining terms and/or classification systems.

The CDD Metadata Standard significantly adds to the freedom of the metadata producer by allowing for either multiple instances of same level data elements or hierarchical nesting of data elements for two of the Main Theme data elements: "Major Theme Description" and "Minor Theme Description". Nesting same kind data elements is one way for a metadata producer to indicate the relationship between data set themes; the data producer can offer minimal or, hopefully, a great amount of detail about a data set as well as providing an indication of its structure to the data consumer. Major Theme Description is "a brief description of a major component (main topic) of the data set." Minor Theme Description is "a brief description of a minor component of the data set."

The standard offers guidance in terms of widely accepted major and minor themes: a domain of approximately 50 major themes and a domain of approximately 180 minor themes are included in the standard. The SCDD compiled the base domains from the "Statistical Abstract of the United States" and cultural geography publications. The domains were then supplemented by surveying Federal agencies for additional areas of importance. Examples from the Major Theme Description domain include Archeology, Communication, Energy, Environment, Parks and Recreation, Population, Science and Technology, and Transportation. Examples from the Minor Theme Description Domain include acreage, amount, climate, endangered species, industry, mortgages, number, price, rural, site, size, voters,

and wildfires. Free text is a member of both domains to allow for extension.

The SCDD feels the final two data elements record significant information. "Computed Value" is a means of indicating data set values are computed values. The metadata producer specifies the components of the computed value and their relationship. For example, a Computed Value may be "persons per square acre". Alternately, the metadata producer may provide a "Scalar". Scalars are mandatory if data set data values are not actual data values or if the scale is not explicitly identified in a Computed Value. A Scalar is "a factor allowing data users to determine actual data values when applied to the data set data values." For example, a Scalar may be "in thousands". An example of a Scalar not being required is if the Computed Value "persons (in thousands) per square acre" has already been provided. Computed Values and/or Scalars are placed following the lowest level Major Theme Description (if not followed by a Minor Theme Description) or following the lowest level Minor Theme Description.

Figure 5 is a comparison of one attempt at using both standard in describing the cultural aspect of the previously introduced agricultural data set.

GRAPHIC

Figure 5. A comparison of metadata describing the topical aspect of a sample data set using both standards.

6. MULTIPLE STANDARDS DON'T NECESSARILY CREATE MULTIPLE PROBLEMS!

Cultural and demographic data metadata may stand alone to describe a topical data set or, because of the commonalties between the standards, may be used as input to additionally produce geospatial metadata for a data set. Part 4 of the proposed CDD Metadata Standard (Part 4: Clearinghouse Spatial Metadata Crosswalk/Profile) offers data producers and users a comparative crosswalk between the two standards to assist in translating either cultural and demographic data metadata to geospatial data metadata or geospatial data metadata to cultural and demographic data metadata. Indeed, a properly designed search engine could perform the conversion from one to the other.

The SCDD feels the similarities and crosswalk between the two standards prevent the availability and use of a second standard from

placing a burden on metadata producers; rather, multiple standards offer multiple opportunities for providing comprehensive metadata to a wider audience of data consumers. And by emphasizing the thematic aspects, it encourages data users to add their data sets to the NII. The Geospatial Metadata Standard serves as the standard for all geospatial data sets. Using the CDD Metadata Standard allows metadata producers the additional ability to address the important topical aspect of their geospatial cultural and demographic data sets. The SCDD envisions the development of toolkits, similar to those available for creating geospatial metadata, to help metadata producers in both the production of cultural and demographic data metadata and the translation of common elements to a geospatial metadata format to reduce redundancy of work effort.

7. CURRENT STATUS

The SCDD's CDD Metadata Standard has progressed through step nine of a twenty-step FGDC Standards Process. Part of this process includes determining whether the standard follows FGDC Guidelines, such as adherence to the A-16 guidelines, consistency with project scope as previously approved by the Standards Working Group, or conflicts with any existing FGDC standards or standards projects, and other related requirements. If approved, the Standards Working Group sends the draft standard to the FGDC Coordination Group with a recommendation that it be announced for public review. If not approved, the specification is returned to the subcommittee for additional work.

To meet the requirements of step ten, the Standards Working Group will review the CDD Metadata Standard for potential conflicts with the existing Geospatial Metadata Standard. The SCDD feels the two standards are not in conflict. However, a larger issue the Standards Working Group will undoubtedly address in reviewing a data category specific metadata standard is the FGDC's overall approach to geospatial metadata and the importance of assisting the majority of metadata producers, who are experts in the subject matter of their data sets but not necessarily the positional accuracy of their data sets.

The introduction, acceptance, and eventual usage of a standard is an arduous process. The SCDD is presenting its CDD Metadata Standard as a means of further meeting the needs of the large community of geospatially referenced cultural and demographic data producers and consumers.

8. THE FUTURE

Cultural and demographic data will be a significant part of the NII. The provision of properly encoded metadata for spatially based and thematic based queries will be the determining factor in how valuable the data are to the people. As good as the Geospatial Metadata Standard and the proposed CDD Metadata Standard are, there are significant portions of information missing. The future must see a better and standardized way to express the statistical and/or positional quality of the data, the expressed relationships in the data, and the accuracy of its representation in the data set. The latter addresses representation that have been rounded, random noise, and so forth.

Future efforts to achieve meaningful and useful data quality standards must include participants from many fields, including but not limited to the users and statistical communities. The standards must not be burdensome to the data producer or obtuse to the data user. They must be relevant to the purpose, i.e. metadata.

9. REVIEW THE STANDARDS

The proposed CDD Metadata Standard may be reviewed online; it is available for viewing on the SCDD Internet home page (<http://www.census.gov/ftp/pub/geo/www/standards/scdd>). A paper copy may be obtained by contacting the Geospatial Research and Standards Staff, Geography Division, Bureau of the Census, Washington, DC 20233 or by telephone (301) 457-1056, Facsimile (301) 457-4710, or Internet (electronic mail) grass@geo.census.gov.

For those who wish to further compare the standards, the Geospatial Metadata Standard may be obtained from the FGDC's Internet site (<ftp://fgdc.er.usgs.gov/pub/metadata/meta6894.txt>). A hard copy may also be obtained by contacting: Federal Geographic Data Committee Secretariat, c/o U.S. Geological Survey, 590 National Center, Reston, VA 22092, telephone (703) 648-5514, Facsimile (703) 648-5755, Internet (electronic mail) gdc@usgs.gov, or Anonymous FTP: [fgdc.er.usgs.gov](ftp://fgdc.er.usgs.gov).

*Content Standards for Digital Geospatial Metadata, June 8, 1994 version.

1 The FGDC is an interagency committee whose member organizations are Federal

departments and agencies having an interest in the production and use of geospatial data.

2 Executive Order 12906, published in the April 13, 1994, edition of the Federal Register, Volume

59, Number 71, pp. 17671 - 17674, is titled "Coordinating Geographic Data Acquisition and

Access: The National Spatial Data Infrastructure."

3 A complete listing of the currently chartered FGDC subcommittees includes: Base

Cartographic, Bathymetric, Cadastral, Cultural and Demographic, Geodetic, Geologic, Ground

Transportation, International Boundaries, Soils, Vegetation, and Wetlands.

4 Metadata are data about the content, quality, condition, and other characteristics of data.