SPATIAL DATA BASES: DEVELOPING A MEANINGFUL INTERNATIONAL FRAMEWORK FOR SPATIALLY REFERENCED CULTURAL AND DEMOGRAPHIC DATA

Leslie Godwin
Geography Division
U.S. Bureau of the Census
Washington, DC 20233-7400
Tel: (301) 457-1056
lgodwin@census.gov

ABSTRACT

Development of a framework for defining, representing, and ultimately exchanging spatial features has naturally focused on physical or political entities in pursuit of the objective of creating and sharing heterogeneous, spatial data bases. Entity data elements, including their definitions, attributes, and relationships, have been developed and refined for physical entities at national levels. The United States has issued both the Spatial Data Transfer Standards and the Content Standards for Digital Spatial Metadata. Similar standards have been issued in other nations.

Although the importance of physical entities remains unquestioned, another aspect of spatial data bases is beginning to receive much needed attention. The heretofore "forgotten component" of spatial data bases is their cultural and demographic component, i.e. their human dimension. Definitions for cultural and demographic data sets are currently being developed in several countries. As with physical entities, the framework for these cultural and demographic components recognizes the importance of data categorization, topics and characteristics classification, and the geographic unit of coverage and temporal component which are required for identification and use of these data.

As the exchange of information through spatial data bases crossed international boundaries and became worldwide, research at the international level initially focused on the application of sharing national concepts of physical entities across nations. The seemingly straightforward task of defining physical entities has proven to be quite complex in the international arena. For example, one nation's concept of a water body proved to be different than another given cross-cultural, cross-linguistic, and cross-disciplinary comparisons (Mark, 1993). Defining cultural and demographic data at the international level will, at the least, be no less daunting a matter. Addressing the broadening of national cultural and demographic data definitions to meet international needs is a timely issue as the development of most national cultural and demographic data frameworks are at an early stage.

This paper provides some guidelines for achieving an international set of standards for describing cultural and demographic data. Since the majority of the decisions taken on behalf of nations or between nations involves components of the human dimension, removing obstacles to sharing cultural and demographic data in an international environment of heterogeneous spatial data bases is imperative. The success of exchanging complete spatial data sets will enhance the abilities of

researchers and decision makers to achieve more equitable and meaningful decisions.

INTRODUCTION

Standards for spatial data sets serve as the basis for understanding, interpreting, and exchanging the data. Standards for spatially referenced data sets typically address the definitions, attributes, and relationships of and between the entities within a spatial context. Sometimes the standards will include an explanation of the theoretical model used to develop the standard. Sometimes the standards will emphasize definitions by offering detailed glossaries of terms related to applications and/or levels of generalization. The entities are the building blocks which provide the framework for constructing higher level entities such as physical features, cultural features, and so forth. The value of a standard rest in how well it describes and conveys the information contained in a data set.

Two kinds of standards for spatially referenced data are being developed, those that describe how to encode the data for exchange and those that describe the content of the data set. The latter is referred to as a metadata standard and is of primary importance here. A number of national standards for spatially referenced, i.e. geographic, data sets are either complete and have been issued or are nearing completion. To analyze and describe the work worldwide is beyond the scope of this paper. Experience indicates that similar problems exist everywhere when it comes to developing geographic standards. Therefore, the work in the U.S. as represented by the standards issued through the Federal Geographic Data Committee (FGDC) and the work in South Africa as represented by the South African National Standard for the Exchange of Digital Geo-Referenced Information (SANSEDGI) will be the primary examples described herein. The FGDC's Spatial Data Transfer Standard (SDTS) and the Content Standards for Digital Geospatial Metadata will the two examples from the U.S.

One very important and diverse data category of spatially referenced data is cultural and demographic data. Cultural and demographic data center on the "human dimension" and cover an extremely wide range of topics. Topics such as agriculture, business, communications, customs, economics, education, the environment, as well as many other aspects of our daily lives are examples. Standards for describing and exchanging cultural and demographic data are at the early stages. Typically, groups interested in cultural and demographic data are using as their starting point existing national standards for geospatial data.

When comparing the characteristics of spatially referenced cultural and demographic data sets with current geospatial standards, two items of importance arise. Their importance is heightened even more when this comparison includes cross-cultural and cross-linguistic considerations. First is the absence of a theoretical model for clarifying the meaning of cultural and demographic data similar to the model generally accepted for other geospatial data. The development of such a model is a necessary foundation to adequately describing the fundamental entities of cultural and demographic data. Second, given the nature of the data, is how to accommodate the lack of precision in terminology used to describe data sets and data components. This imprecision exists between technical fields, even within the same culture, but is made more imprecise when the terms are from different cultures, languages, and so forth. Any cultural and demographic data set standards must allow data producers the freedom to either reference generic definitions or provide specific definitions while giving data users easy access to the definitions, and do this without placing an undue burden on either the data producer or the data user. This paper proposes a model for cultural and demographic data and suggests a framework for developing an international metadata standard for spatially referenced cultural and demographic data.

THE TIMELINESS OF AN INTERNATIONAL STANDARD ON CULTURAL AND DEMOGRAPHIC DATA

Spatial data is a necessary and integral part of an information system and misunderstandings about the spatial framework can have a major impact on the presentation of associated data. However, the importance of data sets to their users lies in a user's ability to display and/or analyze data about topics of concern and interest against a geographic framework. One type which has received little attention, but has a major impact on each of us, is cultural and demographic data. These data center on the "human dimension." Indeed, the majority of the economic, political, and societal decisions made daily are made based upon cultural and demographic data.

Differing perceptions of cultural and demographic data coupled with the fact the data collector, data producer, and data user may believe they have a precise understanding of this data impacts our lives. Cultural and demographic data are a cornerstone to the decision making process used by policy makers at all levels of governments and in all governments. Further, decisions are no longer based on information gathered just within the borders of the country making the decision, but rather, increasingly many decisions are being made that cross international boundaries and effect the citizens of the world.

Data sets from many cultures are accessible given today's computer and communications technology. Internet provides access to an increasing quantity of data via a few keystrokes. Addressing cultural and demographic data standards which meet international needs is a timely issue as data sets become available internationally. The development of most national cultural and demographic data standards are at an early stage, and attention has remained focused on the geographic framework standards and is only now beginning to shift to the importance of spatially referenced cultural and demographic data. The lack of attention this type of data receives is evident by browsing through the Internet with the assistance of the many available search tools. A recent, informal browse achieved only three matches with "human dimensions" (interestingly, all were related to the human dimensions of global environmental change.) While there were many matches to "cultural" and "demographic," none of the information pertained to classifying, clarifying, and exchanging such data. Recognizing national standards for cultural and demographic data are likely to be produced, it is important that they be developed on the best international model possible. An understanding of the international characteristics will help build a foundation for cultural and demographic data that will cross cultural lines.

DEVELOPING A MODEL

A Conceptual Data Model of Geospatial Entities

The U.S. Content Standards for Digital Geospatial Metadata was developed by the FGDC to provide a common set of terminology and definitions for the documentation of geospatial data (FGDC, 1994). Successful creation of this standard was due in part to the independent but cooperative work in many organizations and the many public discussions occurring prior to finalizing the standard. These actions led to a broad concensus on content and a theoretical starting point. This concensus contains an unwritten understanding or perception of the fundamental units of geospatial data. Most persons working with geospatial data agree the fundamental units are the geographic units of points, lines, and areas (and volume or surface under special circumstances). Base features developed from these units form an important framework for such operations as delineating the boundaries of higher level geographic units. Base features may include roads, rivers, pipelines, etc. The exact features that organizations categorize as necessary base features varies and depends on the organizations' requirements for and use of a data base.

Given that the basic building blocks used to construct both the higher geographic units and the base features are the points, lines, and areas (polygons), then there exists a unit of measurement, position. The building blocks are considered as being positioned through the use of commonly accepted coordinate systems (often latitude/longitude). Different organizations have different terminology for the building blocks as well as different coordinate schemes. The U.S. Bureau of the Census refers to them as 0-cells, 1-cells, and 2-cells and uses latitude and longitude as the unit. The U.S. Geological Survey refers to nodes, lines, and areas. A graphic of this fundamental model and the common building blocks is depicted in Figure 1.

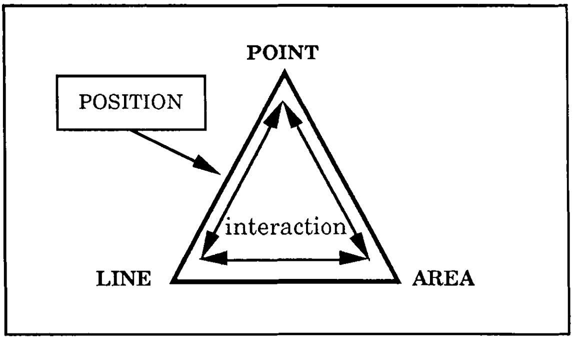

Figure 1. The building blocks of geospatial data.

The Advantages of A Conceptual Model

The fact that common building blocks were accepted did not lead to a single, refined data model nor to a common data structure. Rather, the mechanics of how the data are mentally conceived of in relation to the "real" world and physically organized in terms of structure within a data base varies widely. It also did not lead to a concensus on what constitutes a "correct" classification, or even a "singular," widely accepted set of data definitions for features. An understanding of the building blocks did lead to the additional understanding that there could be different use of the building blocks to create features. For example, roads might be portrayed as lines, perhaps representing road centerlines in one data base, and as polygons in another data base with a width representing say a road's paved surface. Portrayal as road centerlines may be best for the data base user who is concerned with flow or networking data. On the other hand, the data base user interested in siting buildings near highways may need the polygons clearly denoting road widths, right of ways, or some other boundaries to be of the greatest value. The data model provides a common frame of reference for all producers and users of geospatial data. This provides a means by which spatial data can be referenced, even if there is disagreement on terminology assigned to the building blocks.

A Conceptual Model for Cultural and Demographic Data Sets

Whereas for spatial data there exist the fundamental geographic units of points, lines, or areas, a corresponding data unit for cultural and demographic data is elusive at present and difficult to conceptualize. Cultural and demographic data are the data contained in or represented by the geographic framework units. Note, these data and the attempt to clarify them should not be confused with efforts to delineate geographic units which have a cultural aspect as their common denominator.

A number of efforts are underway to delineate culturally-related entities or geographic units. The U.S. military's Tri-Service GIS/Spatial Data Standards (1994 draft) identifies culture as one topic inside the category for delineating geographic units (other topics include landform, geology, soils, hydrography, and climate.) The Tri-Service standard includes identification of such culturally related areas as historical structures, historic maritime sites, prehistoric sites, survey areas, probable and sensitive sites, and native American sites. Although adequate for delineating sites, the standard does not attempt to define, categorize, or classify the cultural and demographic data sets related to these sites.

Just as a fundamental cultural unit has remained vague, a concensus on the building blocks of cultural and demographic data sets has remained equally elusive. Work is currently underway in the U.S. with the purpose of identifying the components of cultural and demographic data. From this effort a conceptual model of the underlying building blocks appears to be emerging.

The identification of the basic cultural and demographic data unit seems to be centered on three questions. Question: Does the data describe a (human) activity (for example an economic activity or a land use activity)? Question: Does the data describe an aspect of humans (for example their health or age)? Question: Does the data describe an aspect of (human) society (for example its political, social, or historical aspects)?

It appears all cultural and demographic data fits into at least one and possibly more of these categories similar to the way in which geospatial data may be categorized as points, lines, and areas. Loosely applying the "geometric" model, one can consider the human as the point, society as the line (linking humans and bounding activities), and activities as taking place over or in relation to an area. And just as there appears to be a common reference to all geospatial features by position, there appears to be a common reference to all cultural and demographic data by count, such as its number or amount. Figure 2 depicts this model graphically.

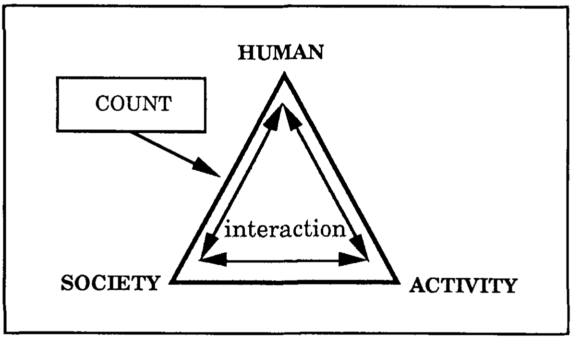

Figure 2. The building blocks of cultural and demographic data.

Though not sufficient for building internationally applicable standards, agreement upon a model such as this can provide a starting point. Once an underlying model is in place, the components for defining the data can be evaluated and various ways of constructing the data proposed and compared. Discussion on the building blocks needs to begin with professional organizations and international associations and conferences providing the forum.

U.S. WORK TOWARD A CULTURAL AND DEMOGRAPHIC METADATA STANDARD

The FGDC's Subcommittee on Cultural and Demographic Data is preparing a standard for describing spatially referenced cultural and demographic data (FGDC, 1994). The Subcommittee is developing this adjunct standard to the geospatial metadata standard because it feels cultural and demographic data are unique from the other more "physically" perceived data and are not adequately represented by using only the geospatial metadata standard. To minimize the time involved in developing metadata, the cultural and demographic data metadata standard is being prepared so that metadata producers and users can easily "crosswalk" between the Digital Geospatial Metadata Standard and the proposed cultural and demographic data metadata standards, as well as the more general Government Information Locator Service (GILS).

The draft cultural and demographic standard identifies a relatively complete description of the contents of a cultural and demographic data set. The principle descriptive parameters in the cultural and demographic data metadata standard are data set identification, themes, geographic framework, temporal framework, source, and data quality. Although all the parameters are needed to obtain a complete picture of the data, the sections on themes, geographic framework, and temporal framework are considered the components most important to cultural and demographic data.

The method for placement of cultural and demographic data sets into themes is the critical operation when applying the standard. Theme prioritization, i.e. nesting, is the most difficult task for the data set producer given the complexity of ideas which can make up a data set. Cultural and demographic data sets often address many major themes and any number of minor themes, all within the same data set. To accommodate a range of combinations of data items and allow the data set producer to prioritize his categories, the Subcommittee developed approximately fifty major themes or categories of cultural and demographic data and approximately two hundred minor themes or categories which can be used hierarchically to further describe the data much in the way attributes are used to describe geospatial features. Although defined in the standards, the major themes are purposefully general and are meant to refer to the general category of the whole data set, even though specific subthemes may be in entirely different categories. Upon determining a major theme, the data set producer may identify as many hierarchies or levels of minor themes as desired depending on complexity and the purpose of the data set.

STANDARD DEFINITION ISSUES

The Geospatial Data Example

Work on spatial data bases has focused on defining the geographic units, the units which provide the base to which additional data may be related. The geographic units may be legal, statistical, natural, or thematic. In the case of legal geographic units, definitions are straightforward and can fairly easily be applied in an international context once the frame of reference is known. For example, if a South African data base is being used by a U.S. researcher the S.A. geographic unit "province" is mentally translated to the geographic unit "state" once the researcher understands a province is the primary legal subdivision of South Africa as a state is the primary legal subdivision of the U.S. The data base user's understanding of the geographic units and their position in the hierarchy in the spatial framework probably easily mirrors that of the data base producer. This argument is valid for the majority of statistical areas.

Although statistical units do not always have boundaries resulting from charters, laws, treaties, or so forth, they are often delineated by governments for data collection and/or tabulation purposes. Their creation for a specific purpose leads to a rather precise definition which can again be translated internationally, though some confusion may arise. If the same South African data base includes enumeration areas, the U.S. researcher might make the mental translation to enumeration district, a low level statistical unit used by the U.S. Bureau of the Census for tabulation. The South African Central Statistics Survey defines an enumeration area as its smallest collection area. In some countries, the basic level of geography for tabulation corresponds to the basic level of geography for collection, consisting of an area one enumerator can cover within a fixed amount of time and containing a limited number of housing units. However, in some countries the basic unit for collection has evolved to being one or a cluster of blocks without regard to number of housing units because the enumeration is conducted electronically. Additionally, in other countries, such as the U.S., both geographic units are used for collection but only the latter for tabulation. The U.S. researcher would have to refer to precise South African definitions to understand that in South Africa the collection unit is also the tabulation unit and probably contains between 200 and 400 housing units[1].

[1] The South African Swuato enumeration area (EA) is one of the exceptions. It is representative of how commonly understood geographic area terms can take on a different meaning to accommodate special situations. Estimates of the Swuato EA population range from one to three million persons and its housing units exceed the planned 200 to 400.

Unfortunately, the simplicity of mental translation becomes more difficult when considering natural or thematic geographic units. Natural or thematic units tend to delineate environmental, physically identifiably similar areas that have boundaries based on the properties or distribution of some variable data. What exactly is a drainage basin? The U.S. researcher may have a definition, but it may not be close enough to another country's definition that the intended use of the data base is not effected. Unlike legal and even statistical geographic units, there are many definitions of a drainage basin and they do not neatly coincide. A drainage basin may be defined as a geographic unit (valley) whose area contributes water to and is drained by a drainage system (one stream and its tributaries). Drainage basins are separated by divides. However, a drainage basin may also be referred to as a watershed; if so, technically the drainage rim is considered to be a part of the watershed (American Geologic Institute, 1976). Inclusion of the drainage rim may, or may not, effect the use of the data base; the total effect depends on two considerations--the extent of the differences in definitions of both the data set producer and data set user and, most importantly, if the data set user is planning to relate additional information to the drainage basin, but use the producer's definition.

Definitions for Cultural and Demographic Data

Definitions, which can be exchanged somewhat successfully about a majority of geographic units, appear to be more complex and elusive in the context of cultural and demographic data. A reason for this is cultural differences must be considered when referencing cultural and demographic data sets, which after all are inherently cultural. Cultural and demographic data sets lack the clarity of their spatial counterparts. Consider as a further component of the South African data base the population count per enumeration area. The U.S. researcher would most likely mentally picture a population count as including a count of the total human population of a given area. The South African data set, however, would not necessarily be clear as to what population of a given area is being reported if the data set were one released prior to 1994.

Definitions may differ uniquely in persons from different cultures due to the cultural bias even in such apparently universal things as time. Just as groups of people or countries' perceptions change over time, individual perceptions also change. A researcher, at age twenty, may include anyone older than fifty in a count of a category of say, "over the hill persons." The same researcher conducting a similar count thirty years later may reevaluate and include anyone older than seventy-five years in the "over-the-hill" count. The resulting data sets and the information which could be gleaned from them would be dissimilar due to changing perceptions, even though the categories stayed the same in title.

There are many examples of varying interpretations of data set items resulting from cultural differences. Cultural misunderstandings are not confined to data analysis; they begin during data collection. Many countries count "households." The U.S. Bureau of the Census considers a household as consisting of a person or a group of persons who make common provision for food and other necessities for living, incorporating a household-housing unit concept (USBC, 1994). In many Moslem communities located in Africa, polygamy is practiced. Multiple wives and their children may live in one compound in a single dwelling, or may live in one compound in multiple closely situated dwellings, or may live in a village in several widely dispersed dwellings. The data collectors may record the counts erroneously based on their cultural interpretation of what constitutes a household. There also may be an error in the population count dependent on whether a husband is counted as spouse to no specific wife, one particular wife, or to several wives. Another example, one effecting data reliability, is age. Generally, persons in some countries know their exact age, perhaps because so many events (beginning school, applying for a driver's license, the right to vote, collecting social security benefits) are based on age. In many other countries people, particularly older people, know their approximate age rather than exact age. The approximation itself may vary, as often the age is approximated around calendars of historical events. The data user may erroneously expects the error associated with age-related data to be as small as the error existent in the data sets more commonly used.

How Standards can Address This Problem

From the FGDC's Subcommittee on Cultural and Demographic Data's experiences certain guidelines for achieving an international set of standards for describing cultural and demographic data are being identified. The first is the need for all those participating in the development of such standards to realize that cultural and demographic data, to a greater extent than other types of data, are susceptible to cross-linguistic, cross-cultural misunderstandings. Because of these cultural differences as well as changes in personal and cultural attitudes which span the temporal dimension, providing a means of referencing a multiple, well-defined, dynamic rather than rigid data definitions becomes extremely important. The task of metadata standards for cultural and demographic data is twofold--they must allow data set producers to easily define their own terms or reference definitions while assuring data set users can easily locate specific definitions.

Many metadata standards allow "free text" entries for terms that are not explicitly defined within the metadata document and allow these terms to be used in domains rather than limiting the data set producer to a closed domain. The freedom "free text" provides to the data set producer is important in accurately describing their data sets. However, with this freedom comes the increased potential for misunderstanding. When "free text" is utilized, either a concise definition of the "free text" or a reference to an easily accessible data dictionary must be assured. Further, data set producers should have the freedom of including within the metadata any additional information they feel to be pertinent to data set use. Standards should not be so rigidly structured that the metadata producers are limited in their ability to provide information they feel is of importance to the data set. This freedom may raise problems in developing a parser for meaningfully accessing the metadata, but the price for overcoming these technical problems are more than compensated for by the value of the resulting increased precision in terminology.

SUMMARY AND RECOMMENDATIONS

This paper did not attempt to build an argument supporting the importance of cultural and demographic data to our daily lives because it was felt to be self evident. Rather, examples demonstrating the need for a standard were presented along with examples of difficulties encountered. A conceptual model of basic units is presented. The model is offered as a starting point only.

If full advantage is to be taken in this age of electronic access to an ever-increasing quantity of cultural and demographic data, a metadata standard is obviously needed. Work must begin now, both within and between countries and cultures. Therefore, the following recommendations are offered:

1. Support research in identifying a robust conceptual model.
2. Undertake cooperative efforts to define and refine a metadata encoding scheme.
3. Make spatially referenced cultural and demographic data sets, their description, availability, and exchange a topic for discussion, presentations, and so forth at both national and international professional meetings and conferences.

It is hoped that the need stated in the paper and the model proposed will encourage development of a truly international metadata standard for spatially referenced cultural and demographic data.

REFERENCES AND ACKNOWKLEDGEMENTS

American Geological Institute, Dictionary of Terms, Revised Edition, 1976, USA.

Broome, Frederick R. and David B. Meixler, 1990, "The TIGER Data Base Structure," Cartography and Geographic Information Systems, 17(1), ACSM, Bethesda, MD-USA, pp.39-48.

Department of Commerce, 1992, 1990 Census of Population and Housing, Summary Population and Housing Characteristics, United States, Department of Commerce, U.S. Bureau of the Census, USA.

Department of Commerce, 1992, Spatial Data Transfer Standard (SDTS) (Federal Information Processing Standard 173), Department of Commerce, National Institute of Standards and Technology, USA.

Federal Geographic Data Committee, 1994. Content Standards for Digital Geospatial Metadata (June 8), Federal Geographic Data Committee, Washington, D.C, USA.

Federal Geographic Data Committee, Subcommittee on Cultural and Demographic Data, 1994, Draft Cultural and Demographic Data Metadata Standards (unpublished), USA.

Lynch, John, September 1994, conversations with Mr. Lynch recorded in meeting notes (unpublished), Central Statistics Service, S.A.

Mark, David M., 1993, "Toward a Theoretical Framework for Geographic Entity Types," prepared for COSIT'93, USA.

(no author), 1994, The Government Information Locator Service (GILS): Report to the Information Infrastructure Task Force (May 2, 1994), USA.

Standards Committee of the Co-ordinating Committee for the National Land Information System, November, 1990, National Standard for the Exchange of Digital Geo-Referenced Information, Version 2, CCNLIS Publication No. 1, S.A.

Tri-Service CADD/GIS Technology Center, 19 November 1993, Tri-Service GIS/Spatial Data Standards (draft for comment), Release 1.2., USA.

The author wishes to thank Mr. Fred Broome and Mr. Leo Dougherty of the U.S. Bureau of the Census for their examples of cultural differences.