UNITED NATIONS SECRETARIAT
Department of Economic and Social Affairs
Statistics Division

ESA/STAT/AC.115/10
May 2007
English only

United Nations Expert Group Meeting on
Contemporary Practices in Census Mapping and
Use of Geographical Information Systems
29 May-1 June 2007
United Nations, New York

Use of Geospatial Technologies for Census Data Collection: Issues and Consideration*

Prepared by

Timothy Trainor
U.S. Census Bureau

* This document is being reproduced without formal editing.

Introduction: Overview of Geographic Support for Data Collection Activities

Geography serves as a foundation for each census or survey. In conducting a census, geography, maps, geospatial data and technology are embedded in most census operations. Statistical data is collected or managed at various levels of geography, for example, at the housing unit level or at an assignment area level, in which technology is the vehicle to design, develop, implement and deploy geographic support. Today, technology supports functions that were completely manual in earlier censuses.

Researching criteria, delineating assignment areas, assigning geographic codes, creating and maintaining an address list, making maps, and updating maps are some of the core geographic activities for data collection in a census.

What part does GIS technology play in each of these activities? The answer to this question depends on many factors, some of which are presented in the issues section of the paper. There is no formula to help answer this question. Each Statistical Agency and Department has its own set of circumstances that contribute to decisions and approaches for a solution.

Components of a Census Geographic Support System

There are three main components necessary to establish an effective census support system: a knowledgeable staff, a dedicated geographic program, and technology to support collection and dissemination activities. An organizational structure comprised of geographers, cartographers, software programmers, database administrators, and information technology (IT) specialists is the mixture of staff in the U.S. Census Bureau's Geography Division. Generally, staff is organized by functional responsibilities into more than twenty working units. The geographers are responsible for determining geographic area criteria and delineation. The cartographers map the areas for both collection and dissemination activities. The software programmers are responsible for writing the code to store and use the data. The database administrators certify the integrity of the data. And the IT specialists keep the hardware running properly to ensure a minimum of downtime. These units work together as a whole to ensure a successful census operation.

Geospatial activities associated with a census require a place to house data elements that are geographically referenced. Examples of a data element include an address that is associated with a housing unit which receives a questionnaire through the postal system, roads to travel to an enumeration area, and boundaries that delimit recognized areas for which data are tabulated. There are many other data elements. A spatial database reflects a data model and organizes and stores the spatial data accordingly.

Software is required to support geographic operations. Applications software is required for a variety of geographic processes such as data capture (digitizing), data processing (query and analysis), and output (mapping). With increasing developments in GIS technology, these basic functions are available in a variety of commercial products. Questions remain however. Do the commercial packages meet each census operational requirement? With the size and complexity of national spatial databases, do the available products provide the processing capabilities in the limited time schedules? Is one product sufficient or are different products necessary?

There are three possible paths for addressing software needs. One option is to develop software in-house while a second alternative is to use commercial software products. A third view is a hybrid approach where commercial products are used and customized, extended, and improved to meet census requirements. Developing a geographic support system entirely in-house is a massive undertaking. At this writing, based on empirical observations, it is unlikely that one "out of the box" product meets all requirements for a census without significant additions and extensions of the GIS software. The hybrid approach is worth noting as commercial products offer a foundation on which to extend and build more and better functional capabilities and production capacity.

To adequately accomplish the goals of a census, a robust IT infrastructure is also needed. Hardware, software, and spatial data are the tools that staff need and expect. Without them, work stops. In addition to these tangible components, one virtual requirement that is critical to successful use of technology is performance. Effective and efficient speed applies to hardware performance, software performance, and Internet communication bandwidth. Inadequate performance is reflected almost immediately in the census schedule.

The Geography Division oftentimes leads the way in hardware advances primarily to support the significant demands of geospatial processing requirements. National spatial databases are very large and improvements in GIS software functionality place huge demands on processors and storage requirements. The historic timeline of hardware began with mainframe computers having sequential processing. The next evolution moved to mini computers with virtual processing that offered greater computing power at the expense of porting and rewriting computer code to work in the new environment. Computer workstations were the next phase with a UNIX operating systems environment, yet another significant shift. Blade server technology is the current approach underway for the next census, where processors are added when needed, that can run different operating systems and different software modules.

Recent Technology Developments Relevant to Census Data Collection

Maturing GIS – In the 1980's GIS was a new concept and was focused principally on spatial analysis. With increasing user demands, GIS has evolved into a robust set of spatial data functions including data capture, geocoding, and mapping. Differences exist in capabilities, functionality, and software and database approaches among the various vendors. The critical question remains. Can GIS today meet the requirements of census data collection activities? The answer to this question lies in a number of factors that are country-specific. Will the GIS product meet census requirements? Does it have the capacity for meeting the production and schedule demands? What level of local support is available when problems occur?

Imagery availability – Expansion of coverage, increases in resolution, and more frequent replacement are recent improvements for access to imagery. The level of spatial accuracy improves while costs are competitive with different types of services. For census organizations, imagery is useful for initial data collection, spatial data updates, and quality control of data provided by other methods. Imagery minimally provides a picture of the feature network at ground level. The amount and quality of viewable data depends on factors such as weather conditions at the time of the image collection and resolution capabilities of the equipment used as well as image processing requirements. For census purposes, the image alone is insufficient. Attributes such as feature classes and street names and addresses, where applicable, oftentimes are minimal requirements. In the absence of other data sources, however, imagery is a good alternative for initial data collection where other options are not available or are more costly. For example, updating national topographic map series on a timeline to meet census schedules is increasingly problematic.

Global positioning systems (GPS) – This available technology is applicable for a variety of census applications. Initial data collection by field listers and enumerators is one use as they carry out data collection activities. This technology is particularly useful in replacing the "sketch map" approach where enumerators draw by sight as they traverse the landscape. Checking the quality of spatial data is another benefit of GPS. In the U.S., the plan for the 2010 decennial census is to capture the location of each housing unit using GPS. This spatial data will be used during a subsequent non-response follow-up operation. In this case, the housing unit location displayed on a handheld GPS-enabled device is used in combination with a mailing address or written house location description to more accurately assure that the enumerator identifies the correct housing unit. There are issues in considering GPS. For example, does the satellite path adequately cover the census area? Are the GPS signals sufficient for the level of accuracy required? Are there enough satellites in the path so that minimal location data is available for an accurate signal? Are there receiving stations for post-processing GPS data to improve quality?

Mapping system development – Census operations require maps to carry out the work. They are used as navigation tools, assignment area descriptions, spatial update instruments, work space for enumerator annotations, and for census operation management planning. Maps depict the areas for which data are collected and tabulated. A large geographic area like the U.S. requires a fully automated mapping approach due to the number of map types, the volume of maps required, and the short term schedules for separate census operations. Use of GIS is a consideration. The same questions posed in the "Maturing GIS" section apply. Minimally, maps have to be functional and usable; otherwise they do not serve their purpose.

Spatial database development – Growth in GIS spawned new interest in geographic data within the database vendor community. Prior to this, some GIS vendors and large organizations such as the BOC developed their own database solutions (TIGER) as the rows and columns of numbers of early relational database models were not the best fit for storing and maintaining topological relationships of points, lines, and areas. This is a technical area that is still developing.

The Evolution of the Census Bureau's TIGER System

The 1980 decennial census was fraught with problems in consistency among and between geographic products. Additions and changes in geocoded address lists, geographic reference files, and maps oftentimes required a similar change to one or both of the other products. These activities were manual and required extensive control. The result was unpredictable inconsistencies. As a result of these problems, a vision was formulated for a single data source that served as the repository for producing all geographic products consistently. Thus began the design of TIGER – Topologically Integrated Geographic Encoding and Referencing.

The most costly and time-consuming aspect of building TIGER in the 1980's was data capture. The U.S. Census Bureau (BOC) collaborated with the U.S. Geological Survey to acquire data layers from their new 1:100,000-scale topographic map series which was designed for scanning technology. The Census Bureau assigned USGS attribute codes to the road network while USGS tagged the hydrography layer. Railroads and miscellaneous transportation features such as pipelines and powerlines (all useful for census boundary delineation) were also digitized.

At the same time, BOC geographers in twelve nationwide regional offices collected feature updates and annotated them on printed 1:24,000-scale topographic quadrangles which were digitized. This was the method used to collect local street names, which generally were not included on topographic quadrangles. In urban centers, the BOC had developed the GBF-DIME files (Geographic Base File-Dual Independent Map Encoding). These were used for address list development, address matching, and address range development for urban areas that had city style mail delivery (house number and street name). These were smaller, limited areas for the 1980 census, compared to the vast expanse of land area in the country. The GBF-DIME files were used as a primary source for TIGER in the urban core and extended out to the nearest 7.5 minute topographic quadrangle. The balance of data was acquired using the 1:100K files provided by the USGS.

The last data type needed was census boundaries. Existing census maps, state department of transportation maps, and special early computer generated plots were used for capturing and inserting the boundary network. Boundaries concluded the list of required data elements. TIGER now had features (roads, etc.), hydrography, boundaries, and addresses.

These separate data elements were vertically integrated to create the TIGER file as a topological file. This meant that there were relationships between each line and area in the database. The software to digitize data from maps, to integrate the separate themes, and to plot source maps and quality control maps was developed in-house. The data model and database structure were programmed in-house. All software applications to insert data and use data following the initial creation of TIGER were developed in-house. As enormous as this task seems, maintaining TIGER over time became increasingly difficult.

Modernizing TIGER in Response to Changing Technology – A Case Study

TIGER was designed to meet the geospatial needs for the 1990 and 2000 decennial censuses. Over that time frame, several significant technology developments occurred. GIS grew and improved. GPS availability expanded. More remotely sensed data were available with new providers and vendors offering new and competitive data and services. Database developers learned the value of geospatial data and began designing technology that addressed geospatial characteristics and uses. Following the 2000 decennial census, plans were initiated to improve the spatial data and modernize the data model, database structure, processing environment and applications that use TIGER data

Access to GPS technology offered an opportunity for the BOC to possibly improve response rates, lower field enumeration costs and reduce the use and management of paper in the large field operations. In the 2010 decennial census, using GPS, enumerators will capture housing unit locations during a nationwide field address canvassing operation where enumerators will verify, correct, add, or delete addresses of housing units where they exist (or don't exist). As part of that task, the enumerator will collect a GPS location using a specially-designed handheld GPS-enabled computer.

Improving TIGER spatial data - In order to assure that the housing unit coordinates fall within the correct census block, it is necessary to improve the coordinate accuracy of the street centerlines, which are the principal features used in delineating census blocks. In 2002, the BOC contracted the Harris Corporation to develop a solution for realigning existing TIGER street centerlines to more accurate locations (to within 7.6 meter accuracy) and to update the features in TIGER using a variety of source materials.

Unlike initial TIGER creation where spatial data was captured and maintained at the national level, with advances and extensive use of GIS, many local, state and tribal governments were acquiring geospatial technologies and were expanding their local capabilities. Through the efforts of the regional office geographers, the BOC maintained an inventory of spatial data at the various levels in a TIGER Enhanced Database (TED). Independent quality checks on the local data were performed to assure it met the minimum quality requirements.

Modernizing the TIGER data model, the database structure, and geospatial processing environment – The original TIGER data structure was based on simple graph theory where topological requirements were assured through use of basic geometries. While this is an approach that returns correct results, there were operational, processing and data user issues. For example, the elements did not correlate directly to real world features and their attributes. Features and boundaries had to be "constructed" from these elements. Where this type of function is needed repeatedly, processing requirements grow significantly. In the modernized version of TIGER, a new data model emulates real world features and provides for efficient expansion when needed.

Another difficulty with the original TIGER was the inability to easily interact with other data sets. The homegrown system did not integrate commercial GIS data from potential partners. In addition, when new data types were determined important for inclusion in TIGER, restructuring the database was a significant effort that had far reaching impacts on most maintenance and data use software.

The design and maintenance of the original TIGER database structure was developed in-house by BOC staff. Maintaining an in-house developed system over a long period of time poses many challenges, particularly as database technology has evolved. For the TIGER modernization effort, the BOC selected Oracle's 10G database technology combined with Oracle's Topology Manager. This combined approach meets the topology requirements and allows for scalability, versioning, and other benefits.

Issues and Considerations in Preparing for Geospatial Technologies for Census Data Collection

The preparation for a census begins with five areas of consideration. First the project proposal needs to be justified. Next the budget must be determined and defended to support the proposal. The geospatial data must either exist or must be collected so work can begin. Lessons learned must be examined to research past successes and failures. Finally the appropriate workforce must be in place to ensure the task can be completed.

Proposal Justification – Investments in geospatial activities are costly and require reasonable and defendable justifications. Answering a few basic questions helps to formulate an approach for the initiative. Why are geospatial technologies needed in a census? What are the benefits? What would be the impact without geospatial technology? Does this investment have benefits to other organizations or to the nation? In pursuing a justification, one view is that it is better to have the case made by an organization outside of the census agency, for example, a professional consultant. If this is a feasible idea, then the work in documenting and writing the justification is done by the consultant with input from census staff.

Funding - Planning for the required funding is paramount in securing all of the components for use of geospatial technology for census taking. Technology, particularly new technology, is costly. Costs for hardware, software, communications, data, and personnel are at a premium with new technology. Cost estimates are not precise, particularly with many unknowns. Overestimating is preferable over underestimating. It is more difficult to get support and additional funding where estimates have fallen short of actual expenses.

Geospatial Data Sources – The key component of implementing geospatial technology is spatial data. What are the sources for acquiring geospatial data? Does a national spatial data infrastructure (NSDI) already exist? Do topographic maps exist that cover enough of the country? Are the topographic maps reasonably up-to-date? Do other government agencies have maps or data such as rural electrification, agriculture, or emergency planning agencies? Are there sources of digital data or maps from local governments, planning organizations, urban centers? Do commercial firms have data and are they interested in partnering? Are there copyright restrictions on the use of the data?

It is unlikely that any two sources of data from data providers will be the same. Data integration is a formidable challenge. No magic button exists that normalizes discreet spatial data sources into a common geospatial database. Current methods oftentimes require significant interactive editing which is costly in time and resources.

Organizational Response to Technology Issues – There are many lessons learned from others who have embarked on the geospatial technology path. Learning about those lessons and how those organizations would approach the task today, based on what they know, what experiences they've had, and exploiting improvements in current technology is time well spent. Should the organization learn and develop its own geospatial capabilities? Doing so allows the organization to maximize the use of the technology with its own resources. If investing in training has the effect of high turnover rates of staff, then other alternatives may be a better option. For example, contracting for a geospatial solution is one approach. Buying expertise must be well managed. Contracting can be expensive. It is important to ensure that contractors deliver what is specified and on time.

In developing TIGER, the BOC did all of the work in-house. Today, census staff works collaboratively with on-site contractors to meet the geospatial requirements. The Geography Division employs contractors for two general purposes: 1) to provide expertise that is not available from existing staff; and 2) to provide additional staff resources on a temporary basis. For example, if a Java programmer is needed and current staff does not have that expertise, then contracted resources are a preferable choice. Learning Java in a class won't meet the demands or schedule of the work. In another example, where additional resources are needed to digitize spatial data, the work may be temporary which is suitable for contracted support.

The Workforce – The BOC Geography Division is comprised of geographers, cartographers, IT specialists, programmers, and database administrators. Approximately half of the Division is comprised of geographers and cartographers and half is comprised of information technology staff. The staff is grouped in Branches according to functions and/or responsibilities. In cases where specific contracted expertise is needed, the contractor works among the staff in the Branch performing that work. In other cases where contracted staff digitize data into TIGER, they work as their own unit on-site at Census Bureau Headquarters.

In one sense, geospatial technologies are enabling technologies for work previously done by geographers and cartographers. In that sense, geographers and cartographers exploit the functionality of the GIS tools. Generally, staff are hired with specializations in a particular aspect of geography or cartography and also have expertise in GIS.

Highly trained specialists such as database administrators for a specific vendor such as Oracle are an important part of the workforce. They are contractors who provide technical guidance and help in managing a large complex spatial database. When doing workforce planning, it is important to estimate what type of staff and how many are needed to meet the requirements. That information is needed for funding estimates and to allow sufficient time to hire highly skilled people.

Conclusion

Geospatial technology has replaced many of the traditional processes formerly done by geography staffs in census operations. While the tools have changed, the type of work has largely remained unchanged. Census geographic area delineation, geocoding, address list development and mapping are examples of functions that are required to adequately and accurately conduct a census. Geospatial technologies have expanded the capabilities and use of geographic support for censuses and surveys. Once in an organization, the technology is likely to spread to other units thus offering greater use of the technology for diverse applications. Like any new endeavor, there are issues and considerations to consider. Learning from others and asking core questions related to how technology improves the geographic support of the census will help in formulating a plan for moving forward.