

User-Driven Integrated Statistical Solutions

Mark E. Wallace and Jonathan Sperling

Abstract: The U.S. Census Bureau and other federal government agencies are exploring how best to respond to increasing user requests for concurrent access to multiple data sets and to integrated data. The Integrated Statistical Solutions (ISS) initiative, under the auspices of the Interagency Council on Statistical Policy, articulates a vision and an historic opportunity for the federal statistical community and the citizens and taxpayers of the United States at the beginning of this next century. Collaborative implementation of the ISS vision will produce a next-generation, customer-driven, cross-program, and cross-agency integrated data access and dissemination service capability available via statistical portals such as FedStats. Activities are currently underway across federal agencies that will provide standards, processes, and tools in the administration of information integration architectures; metadata repositories; product conception, design, and development; statistical visualization; and new disclosure techniques. This paper describes some of the implementation issues and ongoing cross-agency activities to establish ISS. As these efforts begin to succeed, they will help build critical capabilities in the nation's emerging statistical and spatial data infrastructures that will support global, national, regional, local, and individual decision-support systems.

“Government has no more powerful tool to serve customers than information technology. It is central to the Administration's goals of rebuilding the economy and improving the quality of life for Americans.”

National Performance Review

September 1994

“Web technology makes it possible for governments to provide a single point of contact for the public, a single online ‘face’ to structure information according to what is important to the citizen.”

Bill Gates 1999

As we enter this new century amidst a wave of technological innovation, our national statistical community and citizens are being confronted with an urgent and critical need for more accurate, timely, relevant, as well as accessible and interpretable data. More powerful computers and rapidly expanding Internet connectivity have revolutionized data access and dissemination capabilities. There are between 75 and 135 million web users at home and at work today, with a projected growth of 20-25% in the coming years (the Industry Standard 2000). Recent data also point to about 53 million unique users per week (Media Metrix 2000). Federal agencies, in particular, face a loss of credibility and, potentially, support if they fail to deliver the public benefits of this new technology to meet rising customer expectations and fulfill the integrated information needs of our nation and our communities.

The demands of this new social and technological environment on statistical agencies are becoming apparent in other countries as well. Sweden, the Netherlands, Canada, and Australia are among the countries in which statistical agencies are investigating the creation of data and metadata repositories for more strategic use of their statistical data assets, including integrated data products (Sundgren 1998). EuroStat has long been involved in integration and harmonization policies and strategies for national

and regional comparative analysis to support social and economic development in the European Union (van Tuinen et al. 1994). In the U.S., FedStats links more than 70 federal agencies for easy public access to government-produced statistics, but has not yet fully addressed the issue of data integration. In the academic and research communities, a massive Digital Libraries Initiative is being developed to allow easy access, via the Internet, to multimedia information from a diversity of sources.

FedStats, funded and supported by the congressionally mandated Interagency Council on Statistical Policy (ICSP), includes representatives from several federal agencies that are eager to collaborate in the development and implementation of Integrated Statistical Solutions (ISS) (IBM Consulting Group 1999) as an essential engine of the National Science Foundation's Digital Government initiative. ICSP, which includes members from the largest 15 federal statistical offices (Table 1) currently plays and will continue to play a key role in the development and coordination of the nation's emerging national statistical data infrastructure and the emergence of ISS.

Key Statistics	Agency	Department
BEA	Bureau of Economic Analysis	Commerce
BLS	Bureau of Labor Statistics	Labor
BJJS	Bureau of Justice Statistics	Justice
BTS	Bureau of Transportation Statistics	Transportation
Census	Bureau of the Census	Commerce
EIA	Energy Information Administration	Energy
ERS	Economic Research Service	Agriculture
EPA	Environmental Protection Agency	Independent agency
NASS	National Agricultural Statistics Service	Agriculture
NCES	National Center for Education Statistics	Education
NCHS	National Center for Health Statistics	HHS
NSF	National Science Foundation	Independent Agency
OMB	Office of Management and Budget	Executive Office of the President
IRS/SOI	Internal Revenue Service	Treasury
SSA	Social Security Administration	Social Security Administration

Table 1. Agencies represented on the Interagency Council on Statistical Policy

Producing ISS will support the President's Executive Order 12906 (Clinton 1994) by linking statistics to geography and time. Executive Order 12906 initiated the creation of a National Spatial Data Infrastructure, a coordinated effort among federal, state, local, and tribal governments to support public and private sector applications of geospatial data in areas such as transportation, community development, agriculture, emergency response, environmental management, and information. Part of this initiative was to develop a national digital geospatial data framework to support the 2000 Census, a primary statistical benchmark for the nation. The Census Bureau and other FedStats agencies are in a unique position to add significant value to this initiative with a more robust integration of their massive stores of statistical data. At the same time, they must address the very important issues of confidentiality and disclosure avoidance.

Recent surveys, for example, show that two out of three Census Bureau customers use multiple data sets (Dickinson 1996, 1999). These data users would like to dynamically integrate, manipulate, and examine data from multiple sources and time periods to create enhanced content and awareness. Customers accustomed to surfing the Internet are now expecting the capability to make better decisions via ubiquitous access to rich stores of information. As the Census Bureau provides increased on-line access to 1997 Economic Census, Census 2000, and American Community Survey (ACS) data in the near future, expectations will rise even higher. If federal statistical agencies continue to saddle data users with the burden of compiling, sorting, parsing, reformatting, and otherwise putting data from disparate sources into digestible forms before the user is able to use them effectively, they do so at the risk of their own peril. Yet, as we stand at the beginning of the 21st century with the technical and organizational capabilities to address these needs, far too little has been and is being done.

As more and more data from multiple sources and time periods become available, users are increasingly being left with the

additional burden of integrating data sets without the tools or knowledge to know whether the data sets can be integrated or if the results are meaningful. The federal statistical community must work together for the public good to enhance decision-making resources used by our national and local leaders, corporate leaders, small businesses, research organizations, and citizens. Providing data users with the capability to produce ISS via metadata-driven statistical tools that ensure fitness for use will help minimize data user burden and data uncertainty and will maximize data quality and usefulness. Using ISS, a future is envisioned in which all government statistical and geographic data are fully documented according to agreed-upon standards, data confidentiality and privacy concerns are met, and tools and resources are provided to ensure proper usage of the data in an integrated environment.

Perhaps the primary long-term challenge of the ISS is garnering the necessary political, organizational, and financial support. Superordinate cross-agency ventures and information technology infrastructures, such as the ISS, that are not directly tied to the individual missions of the nation's many statistical agencies and departments are difficult to fund. The federal budget system traditionally allocates program funding by individual agency. For these and other reasons, most of the initial ISS work has been based at the U.S. Census Bureau. The Census Bureau is the nation's largest federal statistical agency and primary data collector on detailed population, quality of life, economic, and housing characteristics at all levels of geography.

The Census Bureau has a long history of technological innovation in computers, statistics, and geography (Anderson 1988) and, as the home of the Statistical Abstract for 120 years, an institutional interest in an integrated view of government statistics. In many respects, the ISS builds on the Statistical Abstract concept that has long provided users with a single point of entry into the statistical system. Currently, the Census Bureau maintains the FedStats web server in coordination with other federal statis-

tical and research agencies. Ongoing work discussed in this paper includes wide agency representation and close collaboration with the FedStats community and other public, private, and academic research partners.

The ISS Initiative

The ISS initiative is an effort to retool our data creation and delivery systems to deliver “integrated statistical solutions” to customers inside and outside the federal government. It focuses on delivering what customers want, not advancing particular program areas within the government. The term “integration” is relevant on several dimensions; for example:

- integrating data and metadata (e.g., geography, headers and stubs, definitions, and methodology);
- integrating data across programs and time to make it easier for users to assemble, compare, and analyze time series data; and
- integrating internal product design, creation, and delivery systems to speed access of relevant data to customers.

An “integrated statistical solution” is an answer to a customer’s question, delivered without the customer first having to learn how government programs and/or data files are organized.

Users have long used federal, state, and local statistics to derive answers to their problems, although this has entailed significant work on their part to gather and format the data prior to using it (Cortright and Reamer 1998). But, imagine, if you will, a customer being able to access a federal web site, and to quickly and easily obtain relevant and useful information on whether it makes sense to relocate his/her business and family to Howard County, Maryland based on user-defined parameters. The customer could request data on personally important variables such as retail sales per capita; restaurant sales as a percentage of disposable household income; various age, race, and gender characteristics; the cost of living and crime statistics; and recent student-to-teacher ratios and test scores in public schools. Based on these data, compared to other counties in the Washington/Baltimore (or another) metropolitan area, or perhaps counties nationwide if that is what the customer desires, he/she will have valuable information for use in making a good relocation decision. Or perhaps the user might want to access other available data sets to ascertain other information-based solutions, such as: How is my community doing? How is my family doing? How is my business doing? What sales targets should I set for this county? Where would be the best places to retire? To realize this level of data service will entail collaboration among several local, state, and federal government agencies and their associated databases in the development of an integrated data environment or data web (<http://www.thedataweb.org/>).

Existing data dissemination tools such as American FactFinder (AFF) and FERRETT at the Census Bureau and MapStats at FedStats, although very useful in providing flexible access to individual data sets, are not able to provide a problem-and-solution-oriented view of the data today. By working in conjunction with FedStats agencies as well as state and local agencies,

the academic research community, and the private sector, the ISS initiative is extending these current capabilities. It will provide not only individual products and static data pages, but dynamic interactive tools as well. Currently, it is establishing an interagency collaborative environment for providing dynamically integrated data and metadata that can help provide ISS.

Developing ISS will also address the need to rethink the existing corporate technology infrastructure in most statistical agencies. The current infrastructure is predominantly based on older technology. Many agencies are using outdated systems and are therefore forced to rely on software and hardware that is becoming obsolete. In assessing technical requirements for producing and delivering ISS, “homegrown” solutions need to be replaced by commercially available and supported technology. Rather than pursuing technology solutions in a stove-piped manner, focusing on each agency or business area and its needs, data warehousing, metadata tools, and remote access and integration capabilities will be explored and validated in light of how they can deliver ISS.

The Need For ISS

The Internet has clearly become, for national statistical organizations, the primary distribution channel for data and information. Results of user surveys, focus groups, and ongoing daily contact show that the public is demanding broader and quicker access as well as easy point-and-click retrieval and ordering options. Meeting these demands will lead to vastly larger markets and a growing demand for integrated statistical and geographic information solutions. Similar to the effect that the Census Bureau’s Topologically Integrated Geographic Encoding and Referencing (TIGER) had on the development of a multibillion dollar geographic information system (GIS) industry (Sperling 1995), the ISS program will provide the infrastructure investment that will help spur private sector growth and university research. The ISS initiative could well create an emerging data integration industry.

Efforts by the Census Bureau and collaborative interagencies such as FedStats have traditionally focused on distributing large tabular data sets (e.g., Census Summary Tape File data and public use health data) that contain much more data than many users want. As the focus shifts to information solutions, the user will have access to large numbers of small, focused transactions, giving him/her what is needed within privacy and confidentiality constraints. Rather than opening up a fire hydrant of data, it will allow users to take “sips” of only what they need from multiple data sources and time periods.

Implementation Issues

Data users and providers in the statistical and geographic communities have long been aware of the difficulties in comparing data and/or geography across departments and agencies and/or over time. However, customers have received little support from statistical agencies in their efforts to combine and integrate data from various sources and time periods. As a result, the quality of

individual solutions has been highly variable and plagued with data uncertainty. A primary goal of the ISS initiative is to ensure the thoughtful integration of geographic, demographic, and economic data sets. Achieving this capability within the Census Bureau and with other public agencies will require, among other things, the resolution of significant methodological issues related to definitions of concepts, geography, reference periods, and disclosure avoidance. These issues and challenges resonate across the federal statistical community.

Definitions of Concepts

Data integration is dependent on the use of consistent concepts and definitions across program areas. When programs are administered by different agencies, consistency may be difficult to achieve since program needs may call for different treatments. Many examples, some listed below, exist of where data definitions and concepts differ between censuses, surveys, and administrative records and how they change over time. This situation introduces substantial complexities to the storage, processing, display, and documentation of statistical data in an integrated environment.

Inconsistency and data uncertainty also may result from lack of coordination, documentation, and communication. Data produced by the Census Bureau and other agencies have historically been the responsibility of specific divisions or branches with little communication with other surveys or programs. Documentation, via metadata, of these cross-program and cross-agency “differences” will lead to greater standardization in concepts and definitions. It also will lead to a better understanding of existing and needed differences, a necessary prerequisite for a more robust data integration capability within and between government agencies.

Data definitions, concepts, and units of measure and classification change over time, particularly for social, demographic, and economic data sets. Changes in the way data are defined, classified, collected, and tabulated from one time period to another are often inevitable by-products of attempts to improve the data and/or meet new needs. However, these situations often present comparability problems. For example, many data users have experienced the ongoing issues and effects of changing concepts and definitions of race and ethnicity over time and varying methodologies for reporting these classifications in censuses, surveys, and administrative records. Another example is the recent replacement of the Standard Industrial Classification Code with the North American Industry Classification System (NAICS) and its effect on time series analysis (Zeisset and Wallace 1998).

Methodological differences between program areas and agencies for the same concept also may cause comparability problems. Population characteristics will differ depending on whether they are derived from the Population Estimates Program (administrative records), the Current Population Survey (sample survey), or a decennial census/ACS (universal/sample survey). In the case of business establishments, employment, and payroll, the economic censuses and the County Business Patterns pro-

gram produce data for the same concept, but utilize different methods to arrive at different numbers for the same year.

Using data sets from across the broader federal statistical community can be even more challenging. For example, because estimates of the number of people working in an area are based on different data sources and methodologies, data users derive dramatically different estimates depending on whether they use the Bureau of Labor Statistics (BLS) data from ES-202 Covered Employment, Local Area Unemployment Statistics, or Current Employment Statistics, County Business Patterns from the Census Bureau, or Regional Economic Information System data from the Bureau of Economic Affairs. Hourly wages also will differ depending on whether they are based on data from the BLS Current Employment Statistics, National Compensation Survey, or Occupational Employment Statistics, all of which involve different survey instruments (Cortright and Reamer 1998). ZIP Code data for the 1990 and 2000 censuses were developed from a generalized census block equivalency file, while the economic censuses use actual ZIP Codes from respondents to produce ZIP Code statistics (Sperling 2000).

Further complicating integration issues are instances in which concepts seem to mean the same thing, but are actually different. Demographic data examples include race, ethnicity, or ancestry and measuring social characteristics such as the “digital divide,” disability, or racial discrimination. Residence concepts in censuses, surveys, and administrative records often show differences in definition (e.g., “current” or “usual” residence). Economic data examples include terms such as pay, income, wages, salaries and benefits, and measuring economic characteristics (such as cost of living, inflation, or price change). Differences in confidence intervals from sample data point estimates also pose challenges for data integration and presentation across programs and agencies.

These definitional and methodological issues demonstrate some of the challenges and opportunities to develop more consistent policies and standards along the lines of the work done by EuroStat as well as Statistics Canada in harmonizing data. They also point to the need for providing documentation, including user information, on the “integratability” of various data sets, particularly in our highly decentralized federal statistical community.

Geography

Geography plays a critical role in any future ISS data integration capability. Data integrity and the quality of information solutions are dependent on a full understanding of the relationship between the geographic “containers” of statistical data, the statistical data, and the changes of both over time. Geographic entities used for reporting data often differ across programs and between agencies.

Comparing data from a specific survey or census between two time periods, as well as between different censuses and surveys with different reference dates, are frequent obstacles for robust data integration. Another challenge is comparing data from point-in-time census data with surveys or administrative records

that may use average annual estimates or multiyear rolling averages. The timing of updates to the geographic database (e.g., boundary changes) vis-a-vis the reference period of demographic and economic data collection also presents issues that must be addressed.

Geographic changes and timing of data collection and presentation pose particular obstacles to integrating data from various censuses and surveys. Over time, geographic entities emerge, cease to exist, and change their boundaries either by new technical requirements or definitions, locally requested changes, or legal annexation or detachment. Legal/administrative boundaries, particularly below the county level, are continually changing in many parts of the U.S. Such changes are not within the control of statistical agencies. Until now, most survey data have been available only for very high-level geographies (national, regional, and some states). Generally, many small-area statistical geographic entities (e.g., census tract, block group, and census county division) have been constant from census to census. The advent of the ACS and more timely small-area data available from local agencies may change that paradigm. Further, some geographic entities, such as census blocks and ZIP Code service areas, pose more complex issues because of the ongoing changes in their boundaries and the way that they are defined.

Different surveys use different geographic frames; that is, the universe of available geographic entities for which data are reported may differ from one data set to another. In some cases, the universe of legally existing geographic entities may be modified for statistical reporting purposes. In other cases, some data may not be reported for some or all entities at a specified level of geography. Another situation affecting data integration is the “longitudinal” approach used by some demographic surveys in reporting current data with older versions of geography (e.g., metropolitan areas).

Different reference dates used for collecting and reporting data create uncertainty on the extent of differences between data elements from one data set to another. Demographic and economic data may be collected as of a specified date (i.e., April 1, 2000) or, as in the case of the statistical use of administrative records or the ACS, over a period of time. Even if there are no changes in geography, different reference dates (e.g., legal boundaries for the decennial census refer to January 1, 2000) and/or methodologies to tabulate survey or administrative data for a given time period may pose obstacles to data integration, especially for areas experiencing rapid changes in demographics.

Demographic, economic, and geographic changes occur across time and space. Capturing snapshots over time will always present some obstacles to data integration because of the different reference periods used to collect and present data, the availability of data for specific types of geographic entities across agencies and time, and changes in spatial coverage. These and other temporal and spatial obstacles to data integration are more pronounced at lower levels of geography where change is more dynamic. Solutions to these issues, and the ability to re-use and re-purpose data, will revolve around the further development of

standards, metadata, an enhanced geographic information processing environment, and the ability to re-tabulate data using the respondent location.

Disclosure Avoidance

The protection of inadvertent disclosure of individual persons, households, housing units, establishments, or primary sampling units, especially in public use databases, is a foremost concern of government statistical agencies (Croner et al 1996). While georeferenced identifiers can greatly increase the value of data analysis, they increase the government’s burden to ensure the protection of the individual’s right to confidentiality. Balancing the efforts to maximize the amount of useful statistical with confidentiality requirements is of even more concern in an integrated data environment. As such, the ability to integrate data sets is dependent on the resolution of privacy and confidentiality issues, the development of sophisticated disclosure avoidance methodologies, and newly devised data presentation methods.

Currently, data published in predefined tables that have already passed disclosure avoidance review are suitable for integration. Disclosure problems arise when there are attempts either to make the geography of the different data sources the same or to publish all of the data from the different data sources for the same geography.

While query systems add additional impetus to the longstanding desire to have surveys use comparable geography and definitions, the problem remains that a survey’s geography was often chosen to provide maximum information to its primary users within the usual constraints of sample size and disclosure. Imposing geography driven by another survey is likely to violate at least one of these considerations. A query system that retabulates the underlying survey data with modified geography for comparability purposes would need to guard against disclosure.

In conjunction with research efforts under the auspices of the National Science Foundation’s (NSF) Digital Government Initiative, the ISS will continue to investigate these issues and develop solutions to the extent that is possible. The National Institute of Statistical Science currently has a web site for its digital government project on web-based query systems for disclosure-limited statistical analysis of confidential data. County-level solutions may be easier to implement as it is a reasonably stable geographic unit already common to many surveys and across agencies. Cell suppression and data-swapping technology could offer some solutions to these obstacles. Currently, tests of the AFF query system are being done at the Census Bureau for lower levels of geography using its “dress rehearsal” data with checks for complementary disclosure. Techniques are being developed to minimize the increasing potential for the linkage of records with unique characteristics (e.g., the ability to select multiple race categories from Census 2000 has created 63 possible answers to the census’s race question) by introducing enough noise to protect the data but not so much that it would greatly distort and reduce data quality (Zayatz et al 2000).

Implementation Strategies

The ISS initiative will use and promote a corporate approach to project management within agencies and a collaborative approach across government agencies and the research community as part of the implementation strategy. Working with other agencies and their representatives via the FedStats Product Concepts Working Group, it will evolve from existing dissemination systems within the government through prototyping to demonstrate new concepts and technologies. Those working to produce the ISS are systematically seeking and integrating internal and external user input. They are also developing partnerships with other agencies and the research community in designing, developing, and reviewing prototypes of next-generation data access tools that can be implemented via new and developing technology solutions. Moreover, the centerpiece of ISS will involve addressing the aforementioned implementation issues via the establishment of corporate and federated metadata repositories, data warehouses, standards, business rules and practices, and customer relationship management systems.

Metadata

The implementation of ISS requires the development of a corporate metadata repository. It also requires the development of tools to create, add, and extract metadata from the repository. Metadata, or data about data, are the information describing data content, organization, derivation, and limitations (Gillman and Appel 1994). Without electronic metadata repositories in the various agencies, the ability of ISS to develop integrated information is seriously constrained. Currently, the Census Bureau and the Environmental Protection Agency are well along the way to developing corporate metadata repositories and the BLS is planning to begin work on one shortly.

Documenting data sets is a first step in re-engineering data collection and dissemination and enabling data sharing between agencies. Statistical and geographic databases have been built thus far to support the mandates of single institutions or parts of an institution. All who collect and manage data for activities related to their own responsibilities must need to understand and appreciate the value of those data to others and to collect and structure their data accordingly. To take full advantage of the opportunities offered by the new technologies, business, government, and academia will need to develop, support, and fund metadata on a systematic and ongoing basis as well as promote access for all.

The Internet has clearly changed expectations and heightened knowledge about the ease of access to information as well as broadened the universe of users. Customers, both internal and external, expect technology to provide easy and rapid access to information and data. They want better-documented, usable, and interpretable data. Developing metadata repositories will enable the ability to address the increasing demand for rapid access to data by collecting business information across the full survey/census life cycle. Metadata repository technology, as implemented in a loosely coupled virtual manner, is scaleable. Local repositories will store metadata so that it can be shared and reused on an

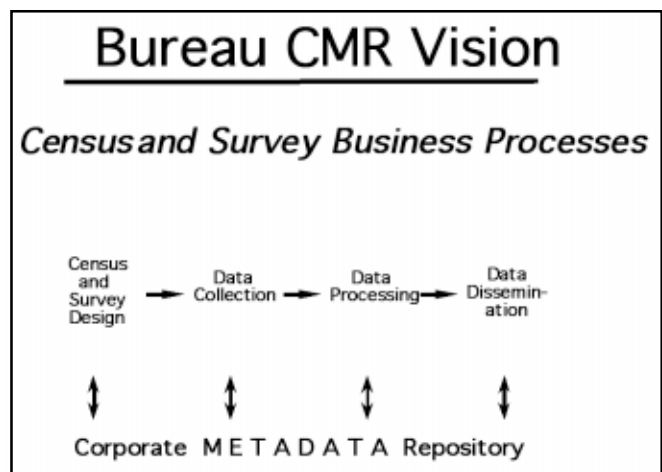


Figure 1. The U.S. Census Bureau's corporate metadata repository vision

enterprise-wide level. These repositories may eventually be expandable to a virtual government-wide (federated) or even global level. Metadata will be the integrating mechanism for providing government-wide information.

Presently, statistical agencies' business processes touch metadata from the survey design through the dissemination life cycle. However, often metadata are not formally cataloged until the dissemination phase when metadata are required for dissemination systems and products. The ISS approach to metadata will be to catalog these data at each phase of the survey life cycle so that metadata can be reused to the greatest degree possible across the various processes. This approach also provides value added during each of the core business processes. Figure 1 depicts the Census Bureau's vision of its corporate metadata repository.

For varying levels of business processes, the repository will be used for creating, updating, and maintaining metadata or using existing systems. Metadata tools can also be used for such purposes as to check additional metadata into the repository, to check out metadata for re-use, or to generate a codebook. If business processes are re-engineered, the new application using metadata can be designed to take advantage of the corporate metadata repository and its functions. Each program area will retain ownership of, and responsibility for, maintaining its metadata. Through the use of tools, the metadata to be stored in the repository are defined and access to the metadata is controlled. Of value to users will be browsing capabilities. In particular, making metadata visible across programs and agencies will allow for re-use and comparison of data sets. The idea is that metadata can be put in once and then used many times.

Supporting Spatial and Statistical Infrastructure

One of the primary assets for integrating government data sets is the TIGER database. The creation of this nationwide digital geographic database by the U.S. Geological Survey and the U.S. Census Bureau for the 1990 census, with linkages to all of the

Census Bureau's extensive social, demographic, and economic data sets, provided a major impetus for GIS developments in the last decade and the ability to easily overlay statistical and other data for common geographic entities. Implementing the ISS can potentially extend these developments and create a robust spatial/statistical data integration industry that will support the National Information Infrastructure and National Spatial Data Infrastructure that will feed national and local decision-support systems.

The Census Bureau's Master Address File (MAF)/TIGER database supports the ISS vision of moving from being an organization of division-specific providers of traditionally separate data products fixed in time and space to being a provider of continuous information solutions using the most current and reliable data and visualization technology available. TIGER/Line, an extract of the TIGER database, is publicly available. The MAF, a nationwide inventory of all residential and, eventually, commercial addresses, is confidential and protected by U.S. Code Title 13. Both files are key components of the nation's data infrastructure that will help enable the linking, integrating, and visualizing of government-wide data sets for statistical purposes. Ongoing updating and enhancement of these files to support census surveys, particularly the ACS and other census operations and programs, will promote even greater use and reliability of TIGER for data integration and visualization.

Improvements in the coordinate accuracy of features, insertion of basic street addresses and their geographic coordinates for all residential and commercial structures into the MAF and TIGER files, and the utilization of Global Positioning System technology for navigation, location, and update activities will create unprecedented opportunities for better handling of data integration across space and time (Sperling and Sharp 1999). Respondent-level coding (e.g., data for a housing unit, group quarters, business establishment, or farm) would permit more precise delineations of statistical geographic areas and user-defined areas and better enable data integration across time and space. However, this capability also raises a number of privacy and confidentiality issues that must be addressed as new systems are implemented.

Interagency Collaborative Projects

"For larger governments, the lesson is to pilot smaller projects to develop expertise and evaluate citizen response." (Gates 1999)

Under the auspices of the ICSP, along with the academic research community, and through the NSF's Digital Government Initiative, various teams composed of staff from a number of federal agencies are researching the potential and validating data integration processes for accessing and integrating both micro- and macro- level data sets (and their metadata). These teams also are developing next-generation data integration tools and products that give customers new data integration functionality and provide information-based solutions not currently available in existing data access tools (e.g., AFF, FERRETT, and MapStats). These interagency teams are addressing many of the aforementioned issues connected with developing ISS.

The FedStats Product Concepts Working Group is currently involved in two major activities: 1) the development of concepts/prototypes for integrated data tools, products, and services, and 2) the development of technical hardware, software, and database concepts/prototypes for integrating data and metadata from multiple sources for producing the aforementioned integrated data tools, products, and services.

Specific activities for the calendar year 2000 include developing integrated data concepts, documenting and using an ever-expanding source of data sets from FedStats agencies and other data sources, conducting usability tests of integrated data input screens and output screens, and defining and refining concepts/prototypes. The current plan is to roll out an initial prototype in Spring 2001 (corresponding with release of Census 2000 data).

In order to facilitate this work, a collaborative work environment with the public, private, academic, and research communities is being set up to test various hardware and software techniques for integrating data from multiple sources. This group will develop database techniques to handle data and metadata from various FedStats agencies to support new integrated data product concepts and to explore the use of real-time, multi-user web tools for integrated data applications.

Conclusion

Converging trends in technologies, databases, customer expectations, and societal needs present the nation's statistical reporting infrastructure with a unique opportunity to achieve a quantum leap of excellence in government service. With its critical mass of data, core competencies, an integrated statistical and geographic processing environment, and a massive data-sharing infrastructure, federal statistical agencies are uniquely positioned to collaborate with local statistical agencies to take advantage of these emerging trends. In the process, they can provide a national leadership role in quality-driven integrated dissemination that will feed many of our nation's critical decision-support systems. In summary, the ISS initiative is good government that makes sense.

Developing this distributed infrastructure and dissemination capability across agencies is an important step in investing in the national information infrastructure that will modernize the delivery of information services and improve data integration tools and products. Providing ISS will enable data users to combine public data resources to maximize use and re-use of our nation's vast data resources.

The ISS presents many challenges and opportunities for the federal statistical establishment. Indeed, cross-agency collaboration and funding on such long-term projects are rare in this age of federal stovepipes and special interest groups. However, there may be no other choice if we are to meet the emerging societal needs for more accurate, timely, and relevant data. While ICSP's support is valuable, there continues to be a need for long-term funding that may be better served by a central funding authority. Through the ISS, research is being conducted on implementation issues, developing prototypes, and establishing multiple partnerships that will ensure more near-term benefits for the Census

Bureau and the government in general, especially with regard to the imminent release of data from Census 2000.

The ability to integrate, overlay, and visualize data and geography is increasingly critical for our national, regional, and local decision-support systems in fields such as transportation, community development, agriculture, emergency response, public health, environmental management, and information technology. The ability to provide current and accurate data that can be quickly integrated with diverse government-wide data sets is critical to the nation's economy and is a means to provide a better quality of life for all American citizens. The ISS initiative is an answer to this call.

About The Authors

Jonathan Sperling is currently a geographer at the U.S. Census Bureau. He has led and worked on various collaborative projects to enhance and better integrate the nation's TIGER data base and Master Address File as well as ongoing bureau-wide and interagency efforts to enhance on-line data access and dissemination services. Dr. Sperling is a contributor to the recently released Congressional Quarterly's "Encyclopedia of the U.S. Census". He received his Ph.D. and Masters degrees from Columbia University where he was a Fulbright-Hays Fellow, Herbert H. Lehman Fellow, and International and Public Affairs Fellow. His current work and research interests include web-based data access and integration, product concepts, data quality, and public health issues. The author may be contacted at the US Census Bureau Geography Division, Washington, DC 20233. (301)457-1100. jsperling@geo.census.gov.

Mark E. Wallace is Chief of the Economic Planning Staff at the U.S. Census Bureau. A graduate of Valparaiso University, he has directed a number of major Economic Census and current surveys programs. For the past five years he has been responsible for directing the Census Bureau's Economic Programs overall data product development, dissemination, metadata, and marketing efforts. Moreover, for the last two years he has directed planning for integrating information access and dissemination on the Census Bureau's web site that allows internal and external users to access and use information drawn from all areas of the Census Bureau and other agencies. He currently is leading an interagency FedStats product concepts team in the planning and implementation of integrated statistical solutions for various user needs. His main areas of expertise are in dissemination of data products and metadata management. The author may be contacted at US Census Bureau, 4700 Silver Hill Rd., Suitland, MD 20746. (301) 457-2621. mark.e.wallace@census.gov

Acknowledgements

The authors would like to recognize all the FedStats team members working on this project under the leadership of the Interagency Council for Statistical Policy. Ongoing research and development work for next-generation information products is being supported by a grant from participating FedStats agencies.

References

- Anderson, M.J., 1988, *The American Census: A Social History* (New Haven, CT: Yale University Press).
- CASRO Inter-Divisional Group, 1997, *Integrated Computing Environment Initiative* (Washington, DC: U.S. Bureau of the Census, Computer Assisted Survey Research Office).
- Clinton, W.J., 1994, *Executive Order 12906, Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure* (Washington, DC: The White House, Office of the Press Secretary).
- Cortright, J. and A. Reamer, 1998, *Socioeconomic Data for Understanding Your Regional Economy: A User's Guide* (Washington DC: Department of Commerce, Economic Development Administration).
- Croner, C., J. Sperling, and F. Broome, 1996, Geographic Information Systems: New Perspectives in Understanding Human Health and Environmental Relationships. *Statistics in Medicine*, 15, 1961-1977.
- Dickinson, J., 1999, *Survey of Census Bureau Current Economic Programmer's Customers* (Washington, DC: U.S. Bureau of the Census).
- Dickinson, J., 1996, *Survey of Users of 1992 Economic and Agriculture Census Data, Final Report* (Washington, DC: U.S. Bureau of the Census).
- Gates, W., 1999, *Business @ the Speed of Thought* (New York, NY: Warner Books, Inc.).
- Gillman, D.W. and M.V. Appel, 1994, Metadata Database Development at the Census Bureau. Presented at the UN/ECE Metadata Workshop, Geneva, Switzerland.
- IBM Consulting Group, 1999, *Integrated Statistical Solutions (ISS) Technology Roadmap* (IBM Consulting Group).
- The Industry Standard (accessed 07/21/00), <http://www.thestandard.com/research/metrics/>
- Sperling, J., 2000, An ISS Proposal for Improving ZIP Code Summary Statistics. *ISS Working Paper*.
- Sperling, J. and S. Sharp, 1999, A Prototype Cooperative Effort to Enhance TIGER. *URISA Journal*, 11(2), 35-42.
- Sperling, J., 1995, Development and Maintenance of the TIGER Data Base: Experiences in Spatial Data Sharing at the U.S. Census Bureau. In Onsrud H. and G. Rushton (Eds.), *Sharing Geographic Information* (Rutgers University Press, Center for Urban Policy Research), 377-396.

- Sundgren, B., 1998, An Information System Architecture for National and International Statistical Organizations. An invited paper prepared for the Conference of European Statistics, Meeting on the Management of Statistical Information Technology, Geneva, Switzerland.
- van Tuinen, H.K., J.W. Altena, and H.C.M. Imben, 1994, Surveys, Registers and Integration in Social Statistics. *Statistical Journal of the United States Economic Commission of Europe*, 321-356.
- Wallace, M.E., C.M. Landman, J. Sperling, and C. Buczinski, 1999, Integrated Statistical Solutions-The Future of Census Bureau Data Access and Dissemination, *American Statistical Association*.
- Zayatz, L., P. Steel, and S. Rowland, Disclosure Limitation for Census 2000, 2000, Contributed paper, UN/ECE Work Session of Methodological Issues Involving the Integration of Statistics and Geography, Switzerland, <http://www.unece.org/stats/documents/2000.04.gis.htm>
- Zeisset, P., and M. Wallace, 1998, How NAICS Will Affect Data Users, <http://www.census.gov/epcd/www/naicsusr.html>

New Knowledge Database

Visit the searchable knowledge database of URISA materials on the URISA website, <http://www.urisa.org/topics.htm>. The database currently includes abstracts and papers (when available) from the three most recent URISA Annual Conferences. URISA is working to add the presentations from other conferences, as well as those from the URISA Journal and other educational publications owned by URISA.