

TheDataWeb: An Innovative Data Framework
Howard Hogan and Cavan Capps
U.S. Census Bureau

Seminar on Innovations in Official Statistics
United Nations, New York
20 February 2009

The U.S. Census Bureau is developing innovative technology that allows issue-analysts to pull together statistical data on a particular topic from a wide range of web-site sources with relative ease and without advanced technological skills. This technology produces new opportunities to turn the vast statistical resources of agencies and organizations into useful information, but it is challenging to expand its coverage.

**SUPPORTING DECISIONMAKERS BY MAKING
STATISTICAL INFORMATION EASILY ACCESSIBLE**

Decision makers and policymakers need to respond to dynamic situations, particularly unexpected major events that require sophisticated access to official data. For example, “Hurricane Katrina,” the “Housing Bubble,” and the “Financial Collapse” required and continue to demand timely access to and advanced manipulation of data. However, obtaining data from multiple statistical agencies is time consuming. Tabulating them in real time, organizing them, and presenting them in a format that is useful to decision makers may take so much time and so many resources that opportunities for efficient and effective action can be lost.

The Need to Change Numbers into Data and Data into Information for Decisions

Numbers must be turned into information that decision makers can use to “connect the dots” of a challenging issue. The number 10.0 is not a datum. The hypothetical statement “The 2005 poverty rate for the U.S. is 10.0%” is data because context has been added. The more detailed (hypothetical) statement “The 2006 poverty rate for the United States as collected by the American Community Survey for the housing unit population is 10.0% (+/- xx %)” provides more context and meaning and thus more information. Approaching the number from a different perspective, the statement “The poverty rate of 10% (+/- xx) for 2007 is significantly lower than the poverty rate of 12% (+/- xx) for 2005” (again hypothetical numbers) amplifies the information value of the data for decision makers. Without context, a number provides little useful meaning.

The Wider Context

One data point is seldom useful. Multiple data presented in a relevant context become increasingly more useful. In order to determine the effectiveness of a policy or to target resources, data must be organized and placed in one or more contexts. For example, to

allocate resources among different areas, data need to be displayed geographically. To determine if a measure associated with an issue is improving, data must be displayed over time. Understanding the demographics, the economic forces, or the health profile for a given place often requires that data be integrated from several different datasets. In the United States, those datasets often come from different organizations.

Dissemination Challenge

How does a statistical agency present the right data with the right context to meet users' actual needs? There is no central database for all statistical agencies in the United States. Organizations that want to integrate data from multiple statistical datasets currently copy data into a third-party database. Given this process, how does a public agency display the most recent and correct data so they can be used appropriately?

Experience has shown that the process of copying and storing data is resource intensive and that data may not be synchronized among copies made at different times, because the latest copy may be the only one that includes all revisions, prompting users to wonder which copy of the "official" data is correct. Rather than maintaining separate copies of the data that are kept consistent manually, an infrastructure that would use technology to obtain the most recent official data from each statistical agency would be less expensive and more accurate. This challenge is being met by the following programs developed by the U.S. Census Bureau.

A THREE-PART APPROACH

The Census Bureau has taken a three-part approach to respond to the challenges outlined above.

- HotReports organize data into usable information for decision makers who are not statistical professionals.
- DataFerrett provides powerful statistical tools for sophisticated analysts to use for their own work.
- TheDataWeb provides a fast and reliable infrastructure to gain access directly to data stored remotely.

HotReports

HotReports use the tools of DataFerrett to draw data from TheDataWeb about specific variables relevant to an issue and a local area and to display that information in graphs, tables, and maps.

HotReports are targeted to policy and decision makers. Local decision makers have limited time and frequently do not have a statistical background. Even decision makers

with technical skills find that they are prevented from using data because of the amount of time required to identify and organize the appropriate data on a particular issue.

HotReports are topically oriented rather than dataset oriented, which is the way that statistical agencies usually present their data. To enable local decision makers to effectively analyze a particular issue, HotReports are created by professional analysts who choose a range of appropriate variables for the topic and design the way to present them – in charts, maps, small tables, or text. Users of HotReports benefit from easy access to a wide variety of related data from many sources, all available in one place.

More specifically, a HotReport is explicitly designed to support decision-making. It is an interactive report that uses Internet Web 2.0 interactive presentation technology to change data into information; it is not a prepackaged data presentation. Text, interactive mapping, time trend charting, and other business graphics are organized so that a non-statistician can see the dynamics of the issue covered by the HotReport.

This guided use of statistical data is possible because the reports are designed by analysts who are knowledgeable about the issues being examined. The policy maker is given a package of appropriate data that has been selected from the myriad different and potentially confusing statistical measures that exist. In other words, a HotReport is organized to give a view of the data over time or in a particular geography that highlights the patterns in the data in an analytically useful way. The decision maker reviews information rather than being faced with undifferentiated, complex datasets.

The ease of use and dynamic quality of HotReports can naturally assist analyses of issues such as regional economic development, emergency preparedness and coordination, public health planning, grant eligibility, community performance indicators, and similar topics.

DataFerrett: An Analytical Tool

DataFerrett is an advanced web-based analytical tool targeted to sophisticated data users. It provides technology to search through thousands of datasets and millions of variables in a way that allows analysts to compare content and sources and relatively quickly select the ones they want to use. It allows advanced statistical tabulations of variables from multiple datasets, which can also be presented as maps or graphs, to illuminate the complexities of an issue.

This analytical tool has the ability to bring the data context along with the numbers. For example, how data may be combined based on geography is known to the system, as are weighting rules and other constraints such as the universe from which the data were collected. Consequently, analysts are prevented from integrating data incorrectly. For example, they cannot combine variables collected only from households with ones about people who live in group quarters, such as nursing homes and college dormitories.

Typical DataFerrett users include federal and state governments, educational institutions, non-profit organizations, and private commercial users. Commercial users typically include the press, consulting firms, and retail, marketing, insurance, financial, and pharmaceutical companies. Because advanced statistical modeling tools are being added to the current descriptive statistical and GIS capabilities of the system, more researchers will benefit from the system in coming years.

TheDataWeb: Infrastructure Engineering

“TheDataWeb” is the software, hardware, and browser engineering that makes the DataFerrett’s tools and HotReports possible.

TheDataWeb is “smart” data-networking software. To normalize data streaming from many different databases, vendors, statistical packages, and data structures, technical specialists transform the underlying structural rules of each dataset (the metadata) from their original formats to that of TheDataWeb. This transformation enables the tools in DataFerrett to operate on, and integrate, information from many databases in a consistent manner. As people who have used data stored in different formats and schemas are aware, this work is a challenging undertaking and it often presents such a major obstacle that highly valuable data integration does not occur. The initial investment in creating consistent documentation for many datasets pays off repeatedly, by enabling this software framework to provide data searching capacity and rules for proper statistical usage and data integration that are invisible to the user.

The beauty of TheDataWeb is that users need to learn only one set of tools to be able to access and manipulate data from a variety of organizations and sources. Perhaps its most important contribution is the ability to integrate information from different sources on related facets of a chosen topic. Hence, this network becomes more useful as more datasets are connected to it.

CHALLENGES FOR THE FUTURE

The Census Bureau collects and disseminates some the most complex U.S. statistical data. The U.S. decennial census, including the American Community Survey, is only one of a wide array of statistical efforts managed at the Census Bureau. They include complex cross-sectional and longitudinal surveys of households and establishments, as well as innovative use of administrative records. Their formats include micro-data, aggregate tables, model-derived estimates, and time series. TheDataWeb team has spent over a decade developing the ability to handle each type in ways that are invisible to the user. As new ways of organizing datasets and more complex metadata formats are developed, they too must be integrated into TheDataWeb to make it as useful to the country as possible.

The way that people look at and use data also evolves as user interfaces develop new formats and displays. We must continue to work with other organizations to stay

current with continually changing user expectations and technical implementations, such as methods of visualizing analytic results. An important goal is to encourage other organizations to contribute code to implement these new tools. A related challenge is to develop processes to assess code from other organizations to be sure that it meets the strict federal IT requirements.

Performance of the system to distribute statistical data must continually be improved so that growing numbers of users will be able to manipulate massive datasets without increasing the time it takes to see the results of their analyses. Moreover, federal IT security requirements are continuing to expand, and the system's security processes must evolve to meet them.

As TheDataWeb and DataFerrett move forward, electronic connections with other agencies' public-use datasets need to expand. Some currently exist, so the Census Bureau can obtain the most current data from those agencies electronically. However, at the moment, many of the databases in TheDataWeb are copies of other agencies' files that reside on the Census Bureau's system. This structure has strong disadvantages because when the responsible agency updates the data, the Census Bureau's copy also needs to be updated, requiring human intervention and opening the opportunity for inadvertent error and delay. It would be far better if the Census Bureau obtained the data electronically, in real time, from files maintained by the responsible agency.

Technology currently exists to allow TheDataWeb to connect electronically to public-use data maintained by other organizations with minimal impact on them. In some cases, the primary organization's system for the public to access the data they have released has sufficiently large capacity that no expansions are needed to incorporate their data into TheDataWeb. In other cases, approaches such as using a mirror server that is automatically updated whenever the data on the primary server are updated could prevent disruptions, such as overloading the primary system and thereby increasing response time for existing users.

TheDataWeb in no way substitutes for the websites that statistical agencies have already developed. Its value is to leverage the accessibility and usefulness of the particular public-use datasets from various agencies that could appropriately be integrated with each other. TheDataWeb enables analysts to integrate data from several agencies for a particular use, such as a map showing unemployment and energy costs for different geographic areas, without having to master the intricacies of finding and using the disparate data sources. This innovative approach to disseminating data provides significant benefit to both the agencies that release official statistics and the analysts and decision makers who rely on them.