

Adventures in Dataland

Business Data Sources

2012 BRASS Program Planning Committee

BRASS Program Planning Committee Members:

Antony Lin, Chair, Irvine Valley College; Chris LeBeau, University of Missouri-Kansas City; Paul Brothers, University of Alabama; Laura B. Carscaddon, Georgia State University; Jason Dewland, Arizona State University; Allison Leaming, Thunderbird School of Global Management; Julia Martin, University of Toledo; Peter McKay, University of Florida; Michael Oppenheim, UCLA; Michael Siciliano, University of Alabama

BRASS Program Committee Authors: Jason Dewland, Business & Economics Librarian/Assistant Professor, University of Mississippi; Chris LeBeau, Assistant Teaching Professor, University of Missouri, School of Information Science & Learning Technologies, and Research and Liaison Librarian, Bloch School of Management, University of Missouri-Kansas City; Julia A. Martin, Business and Economics Librarian, University of Toledo; Peter Z. McKay, Business Librarian, University of Florida; Michael Oppenheim, Collections and Reference Services Librarian, Rosenfeld Management Library, UCLA Anderson Graduate School of Management.

The Business Reference and Services Section (BRASS) 2012 Program, “Adventures in Dataland: Business Data Sources,” presented a panel of four speakers representing academia, government, and the commercial sectors. Three of the speakers focused on the scope and capabilities of particular data products—Esri, *RAND State Statistics*, and *American Factfinder*—while Princeton’s Bobray Bordelon covered a cross-section of valuable data sources. Resources covered are mostly free; some carry a modest fee.

The BRASS panel featured the following speakers:

- Bobray Bordelon is the Princeton University Pliny Fisk Librarian of Economics and Finance. He also heads Data and Statistical Services, directs the Cultural Policy and the Arts National Data Archive, and is an adjunct faculty member at Rutgers University. He has a BS in Finance, an MLIS from Louisiana State University, and an MBA from New Mexico State University. Bordelon chaired BRASS from 2000 to 2001, and he is its 2006 recipient of the Thomson Gale Award for Excellence in Business Librarianship.
- Angela Lee represents Esri, where she is the libraries and museums industry manager. Esri is the world’s largest GIS (geographic information system) software provider. Lee has fifteen years’ experience coordinating Esri’s outreach and marketing activities to libraries, museums, and related educational organizations. Currently, she is working to help these institutions enhance the educational experience, advance scientific research, and operate more efficiently by leveraging the power of GIS and geographic analysis. Before coming to Esri, Lee was a GIS consultant at Technical Associates in Geographic Analysis, and worked as a library assistant in the Borchert Map Library at the University of Minnesota. Lee holds an MA in Geography with an emphasis on GIS from the University of Minnesota.
- Joe Nation is a professor of the Practice of Public Policy at Stanford University. He directs the graduate student practicum in public policy and teaches policy courses on climate change, health care, and California state issues. His current research is focused on public finance and public employee pensions. Nation is principal at his own consulting firm, California Data and Analysis, where his clients include the RAND Corporation. He directs RAND State Statistics, a collection of databases from the nation’s leading think tank. This association led, in 1997/98, to his participation in the creation of the *RAND California* database.¹ Nation earned a BA in Economics, German,

and French from the University of Colorado, and later, a master's in Foreign Service from Georgetown University. While at Georgetown, Nation worked for both Madeleine Albright and Henry Kissinger. (He diplomatically recalled that Secretary Albright "made a mean baloney sandwich"). Nation holds a PhD in Public Policy Analysis from the Pardee RAND Graduate School. He served in the California Assembly where he authored fifty bills that were enacted into law. He subsequently was named Legislator of the Year. Nation was the principal co-author of AB 32, California's Global Warming Solutions Act.

- Jerry Wong currently serves as the information services specialist for the US Census Bureau, Los Angeles Region. Over the course of four decennial censuses in thirty-two years of service, Wong's experience has been extensive in the Census Bureau's twin missions of collecting and disseminating information. Wong has worked with numerous business groups, including the Small Business Administration, and various chambers of commerce to train entrepreneurs in the effective use of census information for site location and market analysis. Wong holds a BA in Political Science and a Master's degree in Social Welfare, both degrees from UCLA. Wong also has served as a Commissioner for the Los Angeles Unified School District and as an advisor to the Los Angeles City Department of Recreation and Parks.

The interest in data among librarians has picked up noticeably in the past ten years as data archives have grown, as more researchers generate and rely on data, and as data become more accessible to users who have lacked the sophisticated skill level to extract it from its source.

As a nation, we are producing data rapidly and massively: "The federal government quadrupled its number of data centers between 1998 and 2010."² The genetics field alone has yielded terabyte upon terabyte of data. We encounter new approaches to research including data mining, data sharing, data warehousing, and data integration, and we also confront new issues of data privacy and data preservation. When Forbes.com runs article after article on "big data," we know data have arrived. While data do not answer all our questions, data can often confirm, validate, or disprove theories, suspicions, and hunches. Data help us study the past and predict the future; data point up trends and underpin planning; data provide a mirror showing us the impact of our actions, or the lack of them.

A somewhat dark tone was set during the program by the warnings of several speakers about the refusal of the House of Representatives to fund a mandatory decennial census and the American Community Survey (ACS). Although the government has increasingly outsourced government publishing and some of its data gathering, there is no privately produced substitute for the ACS. The loss of these surveys would not only rob libraries and the public of access to necessary data, it would deprive our government officials of vital factual information to guide the governance of the country. This attack

on government appropriations for the Census and ACS ironically coincides with an initiative by the Obama administration to launch its "Big Data Research and Development Initiative" whose members include the National Science Foundation, the National Institutes of Health, the Department of Energy and others.³

We stand at the tip of the iceberg in terms of data gathering techniques and extraction. While data inform us about our personal and societal condition and transactions, it will take political will to use the data to create solutions, not just the political will of legislators, but the political will of we the people.

SELECTED SOURCES OF DATA FOR MARKETING AND DECISION MAKING

Bordelon's presentation covered free data archives that could support marketing research and decision making. Bordelon defines microdata as data representing small units, such as individuals or households, while macrodata is summary data. A great deal of free macrodata comes from the federal government and is published under well-known titles such as *The Statistical Abstract of the United States*.⁴ Bordelon made a forcefully telling point that summary statistics would not be available without the supporting micro level data from such surveys as the *American Community Survey*.

Although Bordelon covered a number of resources, he highlighted three major data archives: the *Consumer Expenditure Survey*, the *Cultural Policy and the Arts National Data Archive*, and the *Survey of Consumer Finances*. The resources in this summary of Bordelon's talk are presented by type of sponsor: US government, university, and nongovernmental and intergovernmental organizations.

US GOVERNMENT

The federal government through its departments and agencies is the largest publisher in the world. One of the better known data collections is the *Consumer Expenditure Survey* (Bureau of Labor Statistics, U.S. Census Bureau).⁵ Summary statistics for individual and family purchases are made available for broad categories such as food, housing, apparel, transportation, health care, entertainment, personal care products, reading, education, and tobacco. Micro level data (for sale) for these categories provide minute market data, such as consumer preference for apples versus bananas. Data collection methods vary—telephone interviews for big-ticket items, personal diaries for smaller everyday purchases. Survey data are based on only a few thousand responses from a national pool. Bordelon cautioned about relying heavily on survey data that are weighted to represent a large unit and then applied to small areas that match that profile.

The *Survey of Consumer Finances* (SCF) (Board of Governors of the Federal Reserve, the National Opinion Research

Center of the University of Chicago) is a triennial survey of US families, with an oversampling of wealthy families.⁶ The survey data include families' balance sheets, pensions, income, and demographic characteristics. During the 2008/2009 financial crisis the Federal Reserve conducted "panel" surveys gathering annual data to assess the current condition of families. SCF offers excellent time series for research. A number of international statistics are organized through federal government sites such as FedStats and the International Statistical Agencies' website.⁷

UNIVERSITY DATA ARCHIVES

Universities also provide rich data archives. The speaker highlighted the *Inter-University Consortium for Political and Social Research* (ICPSR) (Institute for Social Research, University of Michigan),⁸ one of the world's largest archives of social science data. Government agencies and individual researchers provide data to the archive. The ICPSR hosts archives on crime, education, healthcare, fertility, minorities, aging, child care, early education, substance abuse, psychiatric epidemiology, and terrorism. This archive enhances data discovery and can satisfy data archiving mandates. Though membership is required to access much of the data, a fair amount is accessible by nonmembers.

Bordelon is director of the Cultural Policy and the Arts National Data Archive (CPANDA) (Princeton University Firestone Library and the Princeton Center for Arts and Cultural Policy Studies).⁹ CPANDA provides data on arts and cultural participation in the United States. As well as offering fairly unique data that surveys high-end consumers, CPANDA also provides access to the Bureau of Labor Statistics' American Time Use Survey (ATUS) which measures the amount of time people spend on activities such as paid work, childcare, volunteering, and socializing.¹⁰ The CPANDA interface enables users to conduct analysis with microdata without a statistical package.

The Roper Center for Public Opinion Research (University of Connecticut) gathers public opinion data for the purpose of addressing societal issues.¹¹ Content ranges from political to economic to social topics; some data date back to the 1930s. Though heavily focused on the U.S., approximately 90 nations also are represented. The *General Social Survey* (National Opinion Research Center (NORC), University of Chicago) takes the "pulse of America" by conducting social, economic, religious, and political research on American society.¹² NORC has used the same core set of questions since 1972. The American National Election Studies (ANES) (Stanford University and the University of Michigan) conducts surveys on voting, public opinion, and political participation.¹³ ANES is able to query the electorate on religious attitudes and characteristics, areas off limits to the U.S. government.

The Integrated Public Use Microdata Series (IPUMS) (University of Minnesota Population Center) offers completely free population censuses for the U.S. and for countries

around the world.¹⁴ Oddly, the Center holds the largest repository of Census and American Community Survey data, claiming to be the place for "data addicts." The microdata in IPUMS are data representing individuals. Some of the census data extend back to 1850. Although researchers will need to use statistical software, there are a number of data samples accessible through a user-friendly cross-tabulation interface.

Bordelon closed his presentation with non-governmental organization (NGO) and inter-governmental organization (IGO) resources including the World Bank,¹⁵ the OECD (Organisation of Economic Co-operation and Development),¹⁶ the IMF (International Monetary Fund),¹⁷ and Eurostat.¹⁸ Much of the data formerly sold by these organizations are now free.

ESRI

Angela Lee enlightened the audience about Esri,¹⁹ the world's largest GIS (geographic information system) software provider. Esri, founded in 1969 as a land planning research group, introduced the first commercially developed statewide GIS for Maryland in 1973. In 2010, Esri released ArcGIS Online and has enjoyed tremendous success helping businesses use geodemographic data—data about people mapped by geography. Location is one of the key factors for a successful business. Esri's business data help businesses with location analysis, site planning, and market research. While businesses are the heaviest users of ArcGIS, libraries can also make use of GIS for their strategic planning and locating new branches in areas most convenient to patrons.

Lee's presentation included explanation of the different features of ArcGIS, its basic functionality, and the tools available online. The main tools that can be used by businesses and libraries are the demographic data, consumer spending, the Tapestry lifestyle segmentations, market potential, retail marketplace, business locations, and summary data. Lee provided a quick lesson on using the Community Analyst and Business Analyst Online, and finished her presentation by providing links to Esri's online training tools.

Business Analyst Online answers two main questions for businesses: what are the characteristics of a given area, and where can I find new areas that meet specific criteria? To begin to answer these questions, Esri draws from its extensive demographic data. Esri demographers reevaluate their data yearly for their annual estimates and five-year projections. The demographics list contains two hundred variables that can be mapped down to a ZIP code or used for a radius search. Business Analyst Online provides a wizard that guides the user through the selection of an area and lets the user choose demographic criteria from a pop-up box. Features such as color highlighting can be easily customized. As the cursor hovers over selected geography, the user can view a summary of that area's demographic information.

Lee demonstrated the graphic mapping of consumer spending in Business Analyst Online. Users may color code

FROM COMMITTEES OF RUSA

a geographic area based on one of 718 product categories for American consumer expenditure. Lee used an example that graphically showed household health care spending for the Los Angeles metro area by ZIP code. The map revealed that spending on healthcare was greatest in Orange County, while the northern area of central Los Angeles spent the least amount on healthcare. This map potentially guides businesses to direct activities to areas with higher consumer health care expenditures.

Esri's Tapestry lifestyle segmentations are comprised of sixty-five market segments that describe consumers based on socioeconomic and demographic characteristics. The sixty-five segments are organized into twelve life mode groups based on lifestyle and lifestage. There are eleven urbanization groups based on population density, size of city, and location in a metropolitan area. To derive these segments Esri uses cluster analysis and selects data from a variety of sources, including the *Census*, Esri's updated demographics, and independent consumer surveys.

Esri tools in *Business Analyst Online* can help predict the market potential of a given geography. Lee chose an example of nonfiction book purchases by ZIP code for the Los Angeles metro area. Results showed that few nonfiction books are purchased in Anaheim, while areas along the beach front had some of the highest rates of purchases.

Business planners might greatly benefit from the Retail Marketplace option that allows them to discover the areas more underserved by local retailers. By comparing retail sales to consumer spending in thirty-one retail categories, the Esri mapping tool identifies locations where there is excess consumer demand for a particular retail segment. Finally, the business locations and business summary mapping tools provide exact street locations for existing businesses which can be helpful for selecting a location or selling to businesses.

Lee finished her presentation by providing links to training for the Business Analyst Online at training.esri.com. The collection of online tutorials is found at video.esri.com. Additionally, there are a number of suggested classroom activities available at www.esri.com/spatiallabs.

RAND STATE STATISTICAL DATABASES

Joe Nation represented RAND and its original flagship database, RAND California, as well as the subsequent products RAND Texas and RAND State Statistics. The RAND Corporation, headquartered in Santa Monica, California, "is a nonprofit institution that helps improve policy and decision making through research and analysis."²⁰ One of the premier U.S. "think tanks" of the postwar era, RAND (the name is a contraction of the term "research and development") was founded in May 1948, when Project RAND—an outgrowth of World War II—separated from the Douglas Aircraft Company to become an independent, nonprofit organization.²¹

There are 221 unique datasets, or, in Nation's term,

"databases," across the RAND database products. RAND State Statistics covers all 50 states, as the name implies; a complete, detailed breakdown of its coverage may be viewed at states-tats.rand.org/stats/statlist_all.html. RAND California contains some 90 additional databases that are not contained in RAND State Statistics; a detailed inventory of its coverage appears at ca.rand.org/stats/statlist.html.

In both RAND State Statistics and RAND California, the seven following main categories of data are provided: Business and Economics; Community; Education; Energy and Environment; Government Finance; Health and Socioeconomics; and Population and Demographics. To illustrate: in RAND California, the latter topic encompasses 26 subject-specific databases, distributed among the sub-categories Vital Statistics, Population Estimates, Population Projections, and Immigration.²²

Among the especially impressive value-added "databases" that RAND provides in its products, Nation noted, are those for detailed export and import commodity statistics, which comprise some 200 million data points. Adding value to its products is very important to RAND—providing statistical categories, or measures, that do not exist in any other products. For example, for the subject of climate change, and carbon dioxide (CO₂) greenhouse gas emissions, RAND calculates and provides a measure for emissions per one million dollars of Gross State Product. By this measure, California is shown to be four times less polluting than is Alaska. California Greenhouse Gas (GHG) emissions from all sources may be found in RAND California; by comparison, "State Greenhouse Gas Emissions," in RAND State Statistics, provides GHG emissions from CO₂ only (for all U.S. states). "GHG emissions per capita" is another calculation that is uniquely found in RAND databases.

RAND California is updated as frequently as every two to three days. Major updates for both the RAND California and RAND State Statistics user interfaces are scheduled to be rolled out in September 2012; the "Adventures in Data-land" audience previewed these by means of screenshots. A "Housing Price and Transaction Statistics Database" is to be one of the newly available components. This database, already available in RAND California, provides the capacity, with just one click, for dollar results to be adjusted for inflation in any given year, ranging from 1970 to 2011 (to date). Along the lines, again, of RAND working to provide data calculations unavailable in any other resources, this new dataset will have the capacity to provide results according to ZIP Codes. For all RAND product constituent "databases," complete information about the originating data sources is always provided.

As a nonprofit organization, RAND is sensitive to pricing; in the last five years, subscription rates have been increased only 3.1 percent. "Special Requests," for unique, "one-off" datasets, are available, at a cost to produce of \$170 an hour. The "Research Publications" link on the *RAND California* homepage links to the full texts of free RAND publications.

THE US CENSUS AND AMERICAN FACTFINDER

Jerry Wong gave a presentation on the U.S. Census and the new American FactFinder.²³ He began the talk by displaying the U.S. Population Clock displaying a running estimate moment-by-moment of the U.S. population, now more than 313 million. The Population Clock was the segue to a discussion about the Decennial Census which is mandated by Article 1, Section 2 of the US Constitution. The first Census was taken in 1790. Decennial Census population data are the basis for Congressional Apportionment and Redistricting.

Census data can be used to answer questions such as: How many people live here? How has the number of people changed? How old or young are the people? What race and ethnicity are they? How well educated are the people? What languages do they speak? How many are single parents? How many households are low-income?

AMERICAN FACTFINDER (AFF)

American FactFinder (AFF) is the US Census Bureau's reborn web-based data application that gives one-stop access to more than 250 billion data points, including 40,000 tables; 1,500 population groups and tribes; 80,000 business and industry codes; and 12 million geographies. Anyone, anywhere in the world can quickly locate and retrieve US demographic and economic data on population, housing, and geographies. FactFinder includes the Decennial Census, the American Community Survey, the Puerto Rico Community Survey, the Economic Census, the Population Estimates Program, and the Annual Economic Surveys. AFF is a potent research tool that employs sophisticated technology for filtering and searching the database and modifying tables including transposing, hiding, sorting, and filtering rows and columns. Help resources for using AFF include Frequently Asked Questions (FAQs), a glossary, and video tutorials (see example in figure 1).

AMERICAN COMMUNITY SURVEY (ACS)

The most significant change in census-taking in recent years has been the substitution of the American Community Survey (ACS) for the Decennial Census Long Form. ACS collects detailed socioeconomic data from about 3.5 million households in the United States and 36,000 in Puerto Rico each year. It is a large, continuous survey sent to 3,000,000 residence addresses each year to sample, rather than count, population and housing characteristics. ACS is the only source of estimates on social and demographic characteristics for small



Figure 1. Population Growth, 1910–2010
Source: United States Census 2010.

areas and small population groups. The data are used by manufacturers and service sector firms to identify the income, education, and occupational skills of local labor markets. Retailers use ACS to understand the characteristics of the neighborhoods in which they locate their stores. Homebuilders and realtors rely on the housing characteristics in planning developments and selling homes. Local communities use ACS to plan for new schools, hospitals, and firehouses. ACS helps determine how more than \$400 billion in government spending for education, healthcare, and other programs is distributed each year.

AMERICAN COMMUNITY SURVEY (ACS) CONTROVERSY

Remarkably, the American Community Survey has been a source of Congressional controversy in 2012. First-term Republican Representative Daniel Webster, who represents the eighth district of Florida (which includes Orlando), introduced an amendment to the Commerce Department's annual budget bill that would eliminate the ACS. "This is a program that intrudes on people's lives, just like the Environmental Protection Agency and the bank regulators." He also objected to the cost, saying, "We're spending \$70 per person to fill this out. That's just not cost effective, especially since in the end this is not a scientific survey. It's a random survey."²⁵ He took to the floor of the House to denounce the survey as "intrusive" and "unconstitutional" because the law authorizing the ACS requires citizens to respond to the survey. The House passed the bill, HR 5326, with the amendment and another provision eliminating the Economic Census. The fate of the legislation now depends on Senate action. Webster's amendment has been condemned both on the left and on the right, and even in the editorial pages of the *Wall Street Journal*.²⁶ Current Census Bureau Director Robert Groves recently wrote that "eliminating them [ACS & Economic Census] halts all the

FROM COMMITTEES OF RUSA

progress to build 21st century statistical tools . . . this bill thus devastates the nation's statistical information about the status of the economy and the larger society."²⁷

References and Notes

1. "Joe Nation," accessed June 28, 2012, publicpolicy.stanford.edu/peopleprofile/2353andsiepr.stanford.edu/system/files/shared/pubpol/cv_nation.pdf.
2. CIO.gov, "FDCCI Public Plan Links," Chief Information Officers Council, accessed Jun 29, 2012, www.cio.gov/pages.cfm/page/FDCCI-Public-Plan-Links.
3. Forbes.com, "The White House is Spending Big Money on Big Data," May 13, 2012, accessed June 29, 2012, www.forbes.com/sites/reuvencohen/2012/05/13/the-white-house-is-spending-big-money-on-big-data.
4. The government withdrew funding for the Statistical Abstract of the United States. The last edition is the 2012 edition.
5. Bureau of Labor Statistics. U.S. Census Bureau, www.bls.gov/cex (accessed June 28, 2012).
6. Survey of Consumer Finances, accessed June 28, 2012, www.federalreserve.gov/econresdata/scf/scfindex.htm.
7. FedStats, accessed June 28, 2012, www.fedstats.gov/regional.html; U.S. Census Bureau, International Statistical Agencies, accessed June 28, 2012, www.census.gov/aboutus/stat_int.html.
8. Inter-University Consortium for Political and Social Research (ICPSR), accessed June 28, 2012, www.icpsr.umich.edu/icpsr-web/ICPSR.
9. The Cultural Policy and the Arts National Data Archive (CPANDA), accessed June 28, 2012, www.CPANDA.org.
10. Bureau of Labor Statistics, American Time Use Statistics, accessed June 28, 2012, www.bls.gov/tus.
11. The Roper Center for Public Opinion Research, accessed June 28, 2012, www.ropercenter.uconn.edu.
12. General Social Survey, National Opinion Research Center and University of Chicago, accessed June 28, 2012, www3.norc.org/GSS+Website.
13. The American National Election Studies, accessed June 28, 2012, www.electionstudies.org.
14. The Integrated Public Use Microdata Series, University of Minnesota Population Center, accessed June 28, 2012, www.ipums.org.
15. World Bank, accessed June 28, 2012, www.worldbank.org.
16. Organisation of Economic Co-operation and Development, accessed June 28, 2012, www.oecd.org.
17. International Monetary Fund, accessed June 28, 2012, www.imf.org.
18. Eurostat, accessed June 28, 2012, epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home.
19. Esri, accessed June 28, 2012, www.esri.org.
20. Rand Corporation, accessed June 28, 2012, www.rand.org.
21. Rand Corporation History, accessed June 28, 2012, www.rand.org/about/history.html.
22. Rand Corporation, accessed June 28, 2012, ca.rand.org/stats/popdemo/popdemo.html.
23. American Factfinder, accessed June 28, 2012, factfinder2.census.gov.
24. Catherine Rampell, "The Beginning of the End of the Census?" *New York Times*, May 19, 2012, accessed June 30, 2012, www.nytimes.com/2012/05/20/sunday-review/the-debate-over-the-american-community-survey.html?smid=pl-share.
25. "The GOP's Census Takers," *Wall Street Journal*, May 12, 2012, accessed June 30, 2012, search.proquest.com/docview/1012534366?accountid=10920 (accessed June 30, 2012).
26. Robert M. Groves, "A Future Without Key Social and Economic Statistics for the Country," *Director's Blog*, US Census Bureau, May 11, 2012, accessed June 30, 2012, directorsblog.blogs.census.gov/2012/05/11/a-future-without-key-social-and-economic-statistics-for-the-country.