Opportunities and Challenges for using Big Data

Richard L. Clayton

David M. Talan

U.S. Bureau of Labor Statistics

24th Meeting of the Wiesbaden Group on Business Registers

Vienna, Austria

15–18 September 2014

Session 9

The views expressed in this paper are solely those of the authors and do not necessarily reflect the official positions or policies of the US Bureau of Labor Statistics or the views of other staff members.
Motivation

The demand for information, insights and data is increasing in every segment of society. The ability to respond to ever increasing demand forces government statistical agencies to search for new opportunities and new methods.

It is well recognized that the ability to quickly change existing surveys or mount new surveys to collect increasing amounts of new data is extremely limited. Adding new questions to existing surveys takes some time, resources and adds to respondent burden. Building new surveys, with high costs, can take several years. The BLS Green Goods and Service Survey (GGS) was funded in FY 2010. The first annual results were published in March 2012 almost 2½ years later.

The monthly Job Openings and Labor Turnover Survey (JOLTS) was funded in late 1998. The first results for February 2004 were published in April 2004, almost five years later. The resulting time series was still very new when the Great Recession began in December 2007; the JOLTS series was too new for us to fully recognize impending economic collapse.

In these days of instantaneous demands, waiting a few years to build new data sources, investing valuable staff to the new development, then waiting for a sufficient time series to build for analysis, is usually an insufficient response to the original demand. If there is any other way to gain sufficient insight into a problem without waiting for several or many years to get a few answers to basic items, data users will gravitate to other, immediately available sources.

Big Data concepts are based in part on using available data in new ways to address existing and evolving information needs.

The Bureau of Labor Statistics’ Quarterly Census of Employment and Wages (QCEW) program produces a dataset that some consider Big Data. There is no standard agreed upon definition of Big Data, but it is often defined as a data set with the following dimensions: volume, velocity, variety, and veracity. While the QCEW does not have the velocity of some other datasets, it certainly does have the volume, variety, and veracity that characterize Big Data.

Objectives

This paper describes the opportunities and challenges for using Big Data at BLS through its Business Register. The first section of this paper presents various definitions of big data and provides some background on the BLS Business Register. The second section describes some opportunities for using big data within BLS and incorporating it into the Business Register. Some potential opportunities include developing new products such as hurricane zone data, employment and wage estimates for non-profits, and employment estimates from foreign direct investment. Section three describes some challenges of using big data that include statistical, legal, and technical infrastructure issues. The last section describes the evolving nature of Big Data and how experimenting with it may yield promising areas of future development.

Section 1. Definitions of Big Data

There is no commonly agreed upon definition of Big Data. Wikipedia describes Big Data as “a blanket term for any collection of data sets so large and complex that it becomes difficult to
process using on-hand data management tools or traditional data processing applications.” (Wikipedia, 2014).

Another definition of big data often cited is the 3 V’s definition: volume, variety, and velocity. This definition was first created in a 2001 research paper by META Group (now Gartner) analyst Douglas Laney where he described the challenges and opportunities of data growth as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). The Gartner group and much of the computer industry use this definition to describe big data.

Others add a 4th V: veracity. Veracity is an indication of data integrity and the ability for an organization to trust the data and be able to confidently use it to make crucial decisions. (Villanova, 2014)

Some statisticians have referred to Big Data as non-sampled data, characterized by the creation of databases from electronic sources, whose primary purpose is something other than statistical inference. Others refer to this as “unstructured data,” referred to data that was created for another purpose, but can be used to create something else (examples include online job vacancies).

Big data can be text as well as data. For example, the text on the Internet grows every day. This text can be mined as Big Data. The data generated from sources such as weather related devices, financial transaction logs, biomedical logs, radio-frequency identification (RFID) readers, and wireless sensor networks, etc. may be considered by some to be examples of Big Data.

Another concept of Big Data is to try to use existing data in new ways; to match existing data to other sources allowing new data products and economic information at low cost and low or no burden. Several of these approaches will be discussed below.

Existing data is not just economic data. Paradata is data about the data that can be used to streamline operations, optimize resource use and burden reduction. Paradata has been used for many years, without the term. For survey operations, and in a simple example, mailings might be organized by postal area to maximize postal discounts. Telephone surveys target businesses only during business hours and call scheduling accounts for time zones. Workload planning also includes call times, number of call attempts before declaring a refusal or shifting strategies to “reluctance” approaches.

Another aspect of adding value to existing data is to improve visualization techniques. Graphing and mapping are longstanding visualization tools. More sophisticated applications increasingly are used to make information understandable in new ways. The challenge is in developing staff with the vision and skills to create suitable outputs for existing and new data.

One of the motivations to direct resources to the Big Data concept is the high cost and relative inflexibility of surveys. Surveys are often developed to focus on a single targeted concept to be measured.

Monthly surveys are by definition short and focused due to the overall goal of gathering and publishing very timely data. The BLS Current Employment Statistics survey has been tightly
honied to provide the earliest measure of business activity measured in employment and hours worked. These data are released only 2 to 3 weeks after the reference period. Any added variables, regardless of usefulness, would surely challenge its highly timely data that are eagerly awaited and closely scrutinized.

Similarly, the BLS Job Openings and Labor Turnover Survey (JOLTS), collects job openings, hires, quits and layoffs with near-CES timeliness. Economists have suggested the usefulness of adding more questions to the JOLTS collection. Adding questions on the duration of vacancies might support insights into the tightness of the labor market. Noting whether a hire is a replacement hire or a new job would provide insights into employers intentions to add to total employment. Adding occupation of hires would be a very useful indicator of occupational demand and also insight into whether there is pressure on wages. Also, it must be considered whether the existing respondent would have easy access to the new data items sought.

However, in both of these surveys, adding questions to gather new information would necessitate changing forms, systems, and possible changes in sample design, etc. While these changes are not insurmountable, they would require new resources that are currently not available.

**Identifiers:**

Business Registers offer some advantages over surveys as platforms for Big Data initiatives. For example, the wide coverage of the BR does not require sample plan changes. BR’s usually include a number of identifiers that can be used to match to other datasets. Tax systems require mandated identifiers that become widely used for various purposes. In the U.S., the Internal Revenue Service (IRS) provides Employer Identification Numbers (EIN) to every business upon request. These EIN’s are required for filing taxes. Other entities also collect the EIN in other collections making the EIN an extremely valuable tool for matching administrative datasets.

**BLS Business Register:**

The BLS business register is the Quarterly Census of Employment and Wages (QCEW). It provides monthly employment and quarterly wages for all businesses with employees covered by unemployment insurance, about 98% of total employment. The UI administrative data is augmented by two additional collections. First, the Annual Refiling Survey (ARS) updates industry, geography, and respondent information on roughly a three-year cycle. The Multiple Worksite Report (MWR) captures employment and wage data for individual establishments under multi-unit businesses, allowing the QCEW to correctly locate establishments in the correct industry and local area. These three sources combined with intense review and editing of the data provides the best and most current universe for sampling and employment benchmarks. The QCEW is published within 6 months of each quarter.

**Four “V’s”:**

For the purposes of this paper related to business registers, the Big Data concept will focus on the four “V’s”.
Volume: Business registers are by definition voluminous covering all eligible business establishments and/or firms. The QCEW holds 9.3 million establishment records per quarter. The entirety of the longitudinally linked dataset contains 770 million records from first quarter 1990 through fourth quarter 2013. This number grows with the economy and time.

Variety: Business registers can cover a large number of data items providing a rich source of analysis on many dimensions.

Velocity: Velocity refers to the frequency and timeliness of the data. The QCEW is produced quarterly and is published within 6 months after the reference period. Few business registers can match these characteristics.

Veracity: Veracity is also a strength of Business Registers. Mandatory reporting with tax agency efforts to insure complete reporting provides a strong basis for accuracy. Mandatory reporting alone does not guarantee accuracy. Post collection editing and review can further improve accuracy. In the case of the QCEW, a range of edits and validating measures, including direct respondent re-contact and supplemental collections, are applied to turn these UI administrative data into highly accurate and reliable economic statistics.

Section 2. BLS Uses of Big Data

While Big Data is a relatively new term, concepts behind it are not new. The QCEW, for example, is the sample universe for all BLS business surveys. Probability sampling insures that survey responses represent the universe. In addition, it is the employment benchmark for CES, OES and OSHS. As the employment benchmark, the huge coverage, mandatory reporting and intense editing of the QCEW provides the ability to borrow the strength and robustness of the QCEW universe for controlling survey results.

Business Demography: One of the first uses of Big Data by BLS was the development of the Business Employment Dynamics (BED) data from the QCEW business register. BED is the BLS term for business demography. The BED data are created by linking individual business register establishment records longitudinally. These records are then tabulated to create aggregate time series for business establishment openings, closings, expansions and contractions. Additional dimensions of these data have been created to include business age and survival data along with firm size data. These data have gained an increased user base allowing economists and data users to examine the role of employment dynamics in the U.S. economy.

The BED series were first published in 2003. Because it was built on the QCEW database, the BED series were constructed to start from 1993, providing an “instant” decade of data for analysis. Any new series built on the QCEW history may be relatively inexpensively developed to cover multiple years for analysis and depth of insight.

The creation of the BED data elements – age, size, birth, etc. – essentially add new variables to the BR – adding additional variety.
Existing and Available Administrative Data:

The easiest way to add Variety is to merge the existing BR with existing administrative datasets. Assuming a single or several common identifiers, matching records can add new data elements. Identifiers include Employer Identification Number (EIN), business name(s), address(s), etc.

Building on BED: BLS is linking survey data back to the BR sample frame. In this way, it might be possible to apply the existing BED variables (births, deaths, age, size, etc.) to, for example, the Occupational Employment Statistics (OES) survey responses. Such efforts may yield studies on how businesses add occupations as they grow and age; providing new insights into occupational demand of new or growing businesses, or occupations being shed as a business declines or dies.

Similarly, assigning age and size measures to the Occupational Safety and Health Survey (OSHS) microdata may illuminate when (at what age or size) businesses invest in improved practices, or whether the business cycle affects such investments in safety and health.

Both of these surveys are huge, with 400,000 establishments (OES) and a universe of injuries and illnesses (OSHS) each year.

Non-Profit Economy: For example, the QCEW matched publicly available list of non-profit businesses. These new research data covering 2007-2012 will be released in September 2014 meeting a longstanding need for current, detailed industry and geographic detail on this sector of the economy. The non-profit sector covers about 9.8% of employment and has higher than average wages making it important to understand the industry and geographic distribution.

Geospatial: Hurricane Zones: The QCEW microdata contain physical location addresses that have been geocoded giving precise points on the earth. Recently, QCEW staff completed a project to overlay existing hurricane evacuation zones (polygons) over the geocoded business locations. Maps and tables have been published on the BLS website (http://www.bls.gov/cew/hurricane_zones/home.htm) showing the number of establishments, employment and wages that are exposed to potential damage prior to hurricanes of varying strengths. Also, special tabulations provided to the National Oceanographic and Atmospheric Administration (NOAA) allow for studies on the potential risks from sea level rise.

Foreign Direct Investment: Other existing datasets include mandatory reports from the Bureau of Economic Analysis’s Survey on Foreign Direct Investment (FDI). A similar effort to match and review matched and unmatched records will yield a research series meeting demands for this type of data by industry and state. FDI data is in demand for many uses including local economic development efforts. It is easier to target foreign investors from specific countries if there are already investments from a country in a U.S. state or region.

A further extension of this work is to link survey data from the Occupational Employment Statistics (OES) survey to the FDI firms to understand the occupational characteristics and wage distributions of jobs at US affiliates of foreign owned businesses. Building on the currency and frequency of the QCEW, these series would have a relatively short lag increasing the utility for current policy evaluation.
With a pilot test getting underway, and making some assumptions about FDI dynamics, it may be possible to create, with limited resources, a short historical series, again building on the linked longitudinal features of the QCEW.

The QCEW staff in the states reviews thousands of records each quarter and re-contact a large proportion of records with significant changes. The results of these contacts are recorded as “comment codes” explaining these movements. Examples of such comments are:

- strikes
- more business/less business
- weather effects

Further analysis of the patterns of existing codes (by size, age, expanding or contracting) may point to business expectations, future hiring or layoff events.

Also, as a part of the QCEW data stream, we learn about whether a business was bought or sold, or a part was involved in such a transaction. We are looking at ways to study patterns in these events to learn more about the dynamics of businesses, employment and wages.

**Existing but not Available Administrative Data:**

The US federal government has a number of administrative data sets that might, if merged or matched with other datasets yield new information at low costs. However, many of these have very restricted access established years ago. The Office of Management and Budget (OMB) is beginning a process of documenting the range of potential uses and benefits of linking the existing array of administrative datasets with an apparent eye towards attempting to change laws and other barriers, paving the way for fully exploiting these opportunities.

The remaining barriers would be staff resources, linking methods and perhaps new techniques.

**Local Hires:** The National Directory of New Hires (NDNH) was constructed to allow the tracking of parents owing a variety of payments to their children. Businesses are required to report information on new hires within a few weeks to state agencies, which is forwarded to the Department of Health and Human Services to allow cross state tracking and reduce avoidance of child payments. Current laws severely restrict access to these new hire data records. If laws were changed to allow data sharing with statistical agencies, there is potential to link these business records to the QCEW, for example, by common identifiers, to create data on hires and wages at low levels of geography and industry, far more detailed than would be possible with a sample expansion to the JOLTS survey. Even if the NDNH was available and local hires data were calculated, there would still be no detailed data on job openings, quits, or layoffs by state or detailed industry.

**Entrepreneurship:** The Internal Revenue Service (IRS) collects vast amounts of information on businesses. These data are also severely restricted to a small set of agencies and specific uses. The Census Bureau is allowed to obtain these data but not BLS. Decades of attempts to change the laws have failed. If BLS were to have the same access as Census, research and related analysis may improve the consistency of US statistics. Also, access to IRS data on the self-
employed, currently not covered in the BLS BR or business surveys, may allow for new studies on entrepreneurship and the business creation process.

**Section 3: Challenges**

The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization (Wikipedia, 2014).

One critical challenge is building the staff which can handle large datasets, manipulate these, merge or link them to others and bring the needed curiosity to continuously seek new insights, tool, techniques and statistical tools to new and evolving situations (Royster, 2013).

Can we learn how to quantify the characteristics of other datasets in ways that give us the confidence to cite results of analysis? If a dataset has the same distributions as another proven source, can we use it? Or is it enough to provide the dataset characteristics and cite results, letting the user make decisions about suitability?

The community of official government statistics gains it credibility based on professionalism, transparency and sound statistics. We have to maintain those features as we move into Big Data. The challenge is to apply good, sound statistics to new data sets. It will take some years of development and research to learn how to describe and justify other methods than those we have relied on for 50 years.

**Section 4. Future Developments using Big Data using the Business Register**

The scope of Big Data applications is being evaluated at BLS and other agencies. Specifically for the QCEW business register, a number of initiatives have been done, are in process or are being evaluated.

First and as listed above, we are linking the QCEW to every available data set as resources allow. As these projects are attempted and completed, lessons are learned, techniques refined and analyses improved. Resources used on these pilot projects are useful in training staff, developing new techniques and products, working with users and growing the number and types of products on the US economy. While for each project listed above, there are known users. However, it is possible that some new products do not find enough customers to justify continued production. Resources will be moved to other opportunities.

The current array of QCEW-based Big Data projects might be considered “easy wins”, however, the insights, linking approaches and other aspects of learning by doing will inform BLS on opportunities, train staff and build infrastructure for future work.

The new era of Big Data is creating a surge of interest in business registers and their specific characteristics and potential to contribute. It is incumbent on us to profile and market our capabilities during this opportunity.

The Fifth “V” Value: The success in building Big Data capabilities, tools, curious staff and new products will determine the Value of our efforts.
References


Quarterly Census of Employment and Wages website:  http://www.bls.gov/cew/home.htm


“What is Big Data?”, Villanova University:  http://www.villanovau.com/resources/bi/what-is-big-data/#.U85_uPbD-mc