U.S. CENSUS BUREAU

Improving the Business Register through Data-Sharing:

Uses and Challenges

Brandy L Yarbrough

11/21/2014

(Prepared for FESAC meeting, December 12, 2014)

Introduction

This paper provides an overview of how the Census Bureau has leveraged its data sharing arrangement with the Bureau of Labor Statistics (BLS) in order to improve its Business Register (BR). The two agencies began exchanging selected data for multi-establishment companies in November of 2012 with the aim of improving each of their independently maintained master business lists and ultimately creating more cohesion between their statistical data products. The primary purpose of the Census Bureau's BR is to provide a complete, unduplicated universe of statistical units that can serve as the master enumeration list for the Economic Census (EC). In addition, the BR functions as the primary sampling frame for various other business statistics programs and it also serves as a direct source for creating important annual publications such as the County Business Patterns (CBP) reports. The accuracy and completeness of these statistical units obviously has a direct impact on the success and quality of the different programs that utilize the BR. This is particularly true with regard to large and complex multi-unit (MU) enterprises as they are generally responsible for the great majority of the production and output in most sectors of the economy. As such, the Census Bureau makes a considerable investment in maintaining these important companies on the BR. While the annual Company Organization Survey (COS) has been the primary tool used for this purpose, it typically only covers about 40,000 of the 170,000 active MU companies on the BR. It has been suggested that the BLS data could serve as an excellent, non-survey source for statistical unit maintenance on the BR. In fact, most small and medium MU companies are usually not subject to a comprehensive update on the BR until the five-year EC is conducted. Allowing more frequent updates to these types of companies is one area in particular in which the BLS data has been proposed as being potentially beneficial.

The paper is divided into two main parts. **Part 1** covers the application and use of the BLS data in efforts to improve the Census Bureau's BR. **Part 2** outlines some of the challenges that have been encountered with using the BLS data in an operational capacity.

Part 1—Uses

To date, most of the uses of the BLS data files could best be described as "conservative" or "incremental" in nature. In general, there are have been five efforts in leveraging the BLS data to improve the BR:

1. Adding clients of Professional Employer Organizations (PEOs) to the employer business universe.
2. Improved distribution of payroll tax data for MU companies that are either not covered by the COS or that did not respond to the survey.
3. Better targeting of BR single units that are actually operating at multiple locations. In Census Bureau vernacular these are known as "splitters" since, ultimately, the single establishment will be "split" into two or more locations and become an MU company.

4. Improved industrial classification codes (NAICS) for selected establishments, particularly those that are sub-units of consolidated reporters in the EC.
5. Conducting ad hoc research for selected companies and updating the BR using existing interactive tools.

Each of these efforts are described below in more detail.

PEO Clients:

The paper *Co-employment and the Business Register: Impact and Solutions*[1] provides an in-depth look at how the BR operational model is impacted by co-employment arrangements. As such, that information is not repeated is this document. It is suggested that readers who are interested in more details on this topic consult the *Co-employment…* paper directly.

The basic point of *Co-employment…* is that federal payroll tax data serves as the foundation of the BR and therefore some companies engaged in "non-traditional" forms of employment may not be properly counted in the employer universe. Previous efforts by the Census Bureau to address this issue through their own means have yielded, at best, mixed results and have also proven costly as well as burdensome to respondents.

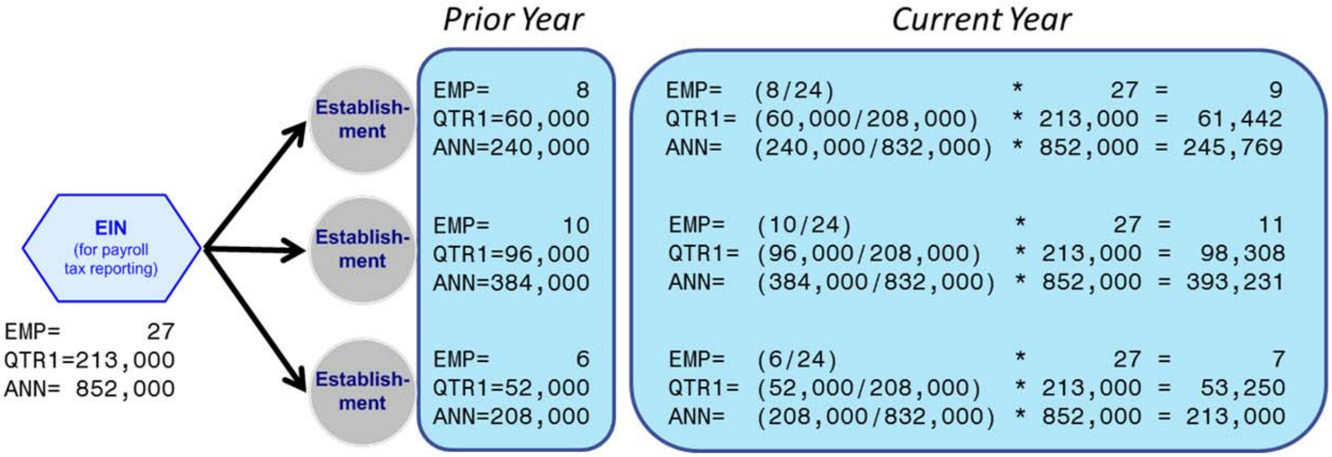
One of the BLS data files provided through the data exchange gives the link between PEOs and their clients. Using both the 2011 and 2012 BLS PEO files, BR staff have been able to add nearly 2,600 of these clients to the employer business universe. These PEO clients account for nearly 130,000 employees and about $9 billion in annual payroll. Most importantly, the largest of these PEO clients were mailed questionnaires in the 2012 EC so that their output measures could be captured and included in published aggregates. Overall, these businesses contributed more than $40 billion in sales, receipts, revenue, or shipments with their most significant presence being in the wholesale sector. It should also be noted that adding these PEO clients to the employer universe on the BR automatically removed them from non-employer status.

Although the BLS PEO client list has helped the Census Bureau improve its coverage of output measures and clarify employer vs. non-employer status, one issue that remains unresolved is the "double counting" of payroll and employment that occurs in the various client industries and in the PEO industry itself (NAICS 561330). One difficulty here is that many of the EINs of the BLS PEOs are actually single units (SU) on the BR. In the current BR operational model, SUs are directly linked to the payroll tax data which is not generally correctable. In order to implement an employment correction at the micro data level these cases would either need to be artificially converted to MU status or a new conceptual statistical unit would need to be developed.

[1] The link http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.42/2013/USA_-_Census_Bureau_Co-employment_and_the_BR.pdf provides access to a copy of this paper.

Distributing Payroll Tax Data

The BR essentially operates on an annual cycle which typically “closes out” in late September or early October in conjunction with wrapping up the COS. One important function that is performed during this close out period is the allocation of federal EIN payroll tax data to the individual establishments of MU companies. This process, which essentially involves distributing current year tax data in proportion with prior year establishment values, is done for all MU companies that are either not included in the COS or that did not respond to the survey. The diagram below illustrates the basic principle using a simple, hypothetical three-establishment MU company:


Prior Year
Current Year
EIN
(for payroll tax reporting)
EMP= 27
QTR1=213,000
ANN= 852,000
Establish-ment
EMP= 8
QTR1=60,000
ANN=240,000
EMP= (8/24) * 27 = 9
QTR1= (60,000/208,000) * 213,000 = 61,442
ANN= (240,000/832,000) * 852,000 = 245,769
Establish-ment
EMP= 10
QTR1=96,000
ANN=384,000
EMP= (10/24) * 27 = 11
QTR1= (96,000/208,000) * 213,000 = 98,308
ANN= (384,000/832,000) * 852,000 = 393,231
Establish-ment
EMP= 6
QTR1=52,000
ANN=208,000
EMP= (6/24) * 27 = 7
QTR1= (52,000/208,000) * 213,000 = 53,250
ANN= (208,000/832,000) * 852,000 = 213,000

This tax data allocation is important not only for maintaining current establishment measures of size but also for providing a complete base of source data that can be used by programs like County Business Patterns (CBP). Unfortunately, since the basis of allocation is prior year proportions, companies that have not reported in the COS or EC in several years may not have accurate distributions, particularly if their business models have significantly changed.

In general, the BLS payroll and employment items found in the provided data files are assumed to be of “equivalent quality” to corresponding values reported on Census Bureau surveys. The basic premise of this application then is, where appropriate, “plug in” the BLS employment and payroll values for prior year imputes on the BR. In theory, this should allow for a more accurate spread of the current year tax data across the establishments.

This process was attempted for the first time during the close out of the 2013 COS. The following criteria were used to determine the universe of candidate companies:

- Must have had no response of any kind or not have been mailed in the 2013 COS
- Must have been totally imputed in the 2012 EC
- All locations must match the BLS MU file on EIN x address
- Payroll tax data must be within 5% of the BLS data at the EIN level

Unfortunately, the results were not very productive since only a few companies covering about 250 establishments met the above criteria. Given this limited number, no action was taken to update the prior year values for the 2013 survey year.

One reason for the low success rate in candidate identification might be that the "prior year" in this instance was the 2012 EC. By design, all MU companies are canvassed in the EC. In contrast, during non-EC years the universe of fully-imputed companies will be much larger since only around 40,000 MU enterprises are covered by the COS. However, the low match rate on EIN x address undoubtedly had a great deal to do with the small number of units identified as well. (In **Part 2**, this matching issue is discussed further).

Targeting Splitters

The EC represents the Census Bureau's most comprehensive effort to identify new MU companies as nearly 2 million single unit (SU) businesses are presented with the opportunity to report that they are actually operating at multiple locations (i.e., "splitters"). Unfortunately, in between censuses, budgets and resources preclude doing anything close to this magnitude. However, in recent years, a small number of SUs-- about 5,000 annually-- have been selected for inclusion in the COS when tax and other administrative data suggest that they may be exhibiting characteristics of multi-establishment companies. Ideally, given that it is a limited and targeted effort, the great majority of these would ultimately become MU companies on the BR. Unfortunately, however, the success rate in converting them from SU to MU on the BR is often less than 40%. Even using the multi-establishment indicator code (MEEI)-- which the BLS routinely attaches to SU EINs that are sent to them quarterly by the Census Bureau for NAICS coding[2]-- has failed to substantially improve this success rate.

Many of the EINs that are on the BLS MU file appear as SU on the BR. As such, it stands to reason that the file might be useful in identifying which SUs would be the most productive to mail in the COS. For the 2014 survey year, a small sample of 100 cases was selected for inclusion in the COS. These cases were on the BLS file as "MU" but on the BR as "SU" and were not mailed in the 2012 EC. Their success rate in conversion from SU to MU will be tracked and compared to other cases in the sample. If warranted, a larger sample of such cases will be selected in the ensuing years.

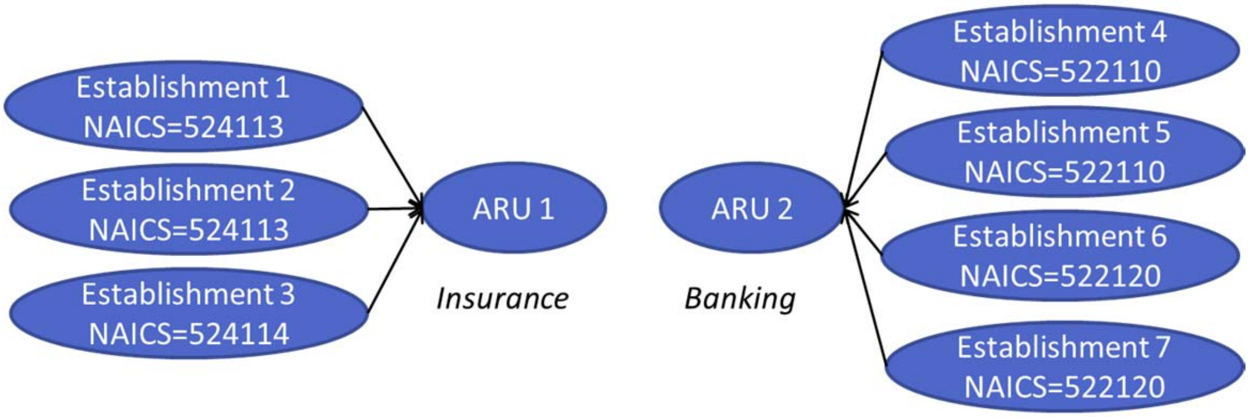
NAICS Codes for Establishments of Consolidated Reporters

In March, 2014 representatives from the Census Bureau and BLS met to review and attempt to adjudicate 40 example cases where the two agencies had assigned different NAICS codes. One interesting trend that was discovered in the review was that several of the cases were establishments that are part of what the Census Bureau calls "alternate reporting units" or ARUs.

[2] This is separate project from the MU data-sharing arrangement. Quarterly, the Census Bureau sends a file of SU EINs to BLS for matching to their business list. The primary purpose is to obtain higher-quality NAICS classifications for these units which typically have either partially assigned or lower-reliability codes on the BR.

An ARU is essentially an aggregation of multiple establishments within a company that are ostensibly operating in the same industry or group of industries.

The diagrams below shows a couple of examples of ARUs:



For most industries that are in scope to the EC, companies are expected to report for each individual establishment. These response data, which include details about specific products, goods, and services of the establishment, are used to verify or update industrial classifications. In many cases, the NAICS code assigned to an establishment is essentially "derived" from its responses to these detailed items via a set of complex algorithms. In lieu of establishment-based reporting in industries where it is difficult or impractical for companies to report detailed output by location the Census Bureau offers ARUs. Some other examples, in addition to the two provided above, include nationwide mining and wireless telecommunications. Although useful for reducing respondent burden and encouraging response to the EC, a drawback is that once an establishment is part of an ARU its NAICS code is unlikely to change unless the respondent explicitly provides a new business description for the location[3]. This is not the most effective mechanism on which to rely as respondents have little motivation to move an establishment out of an ARU industry only open themselves up to even more government reporting.

The question then is: Are there a significant number of ARU establishments on the BR with out-of-date or incorrect NAICS codes? The BLS file provides a potential quality source for verifying and, if necessary, updating these codes on the BR. The general plan will be to review the largest cases after the final 2012 EC NAICS codes have been fed back to the BR. This will allow for changes identified through the EC review to be properly reflected on the BR.

It should also be noted that the BLS file can be used in a similar way to verify or update NAICS codes for establishments that are out-of-scope to the EC but that are in scope to CBP. (The scope for CBP is typically broader than that of the EC). While the COS includes these types of

[3] Although the collection of output measures are consolidated at the ARU level, the respondent is still asked to provide the following by location: names, addresses, EINs, payroll, employment, operational status, and changes in major activity.

establishments, they have the same limitations as ARU establishments in the EC when it comes to capturing changes in activity.

Ad Hoc Research

Specific company-level research is the area in which the BLS data has seen the most pervasive use at the Census Bureau. Making the data available to individual BR analysts as a research tool has helped to improve the unit coverage and data quality of several large companies on a case-by-case basis. In particular, there are some companies that have perpetually refused to provide updated lists of operating locations in the EC or COS. Any lists of establishments that are currently on the BR for these companies have become out-of-date or are significantly incomplete. For some of these companies, the BLS data has been useful in providing more recent updates to these establishment lists. In addition, the payroll and employment data provided in the BLS file can sometimes be used to explain variances between the payroll tax data and survey response data. Although the payroll tax data are generally of high quality there are occasionally errors encountered which can cause a discrepancy when it is compared to the corresponding establishment data as reported on the COS. As such, the BLS data file can serve as another useful source for validating these items. Further the BLS file can be used to verify and, if necessary, adjust the business names, addresses, and NAICS codes of new locations that have been recently added to the BR. Although the COS requests these data for new locations they are sometime poorly reported or incomplete (especially with regard to NAICS codes).

Part 2—Challenges

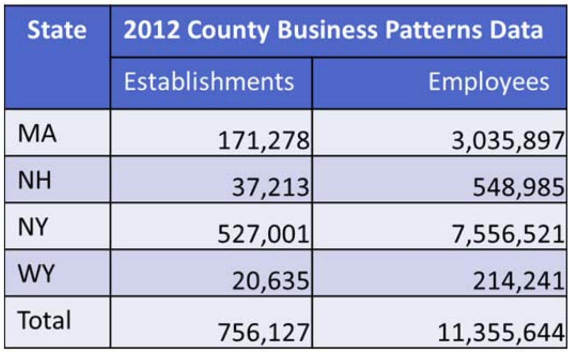
As described in **Part 1**, the BLS MU data have proven to be useful for incrementally improving the BR in a number of ways. However, most of these applications are not what would be called "operational" in the way that, say, the payroll tax data are used in BR maintenance. That is, there are highly defined, regular processes in place that allow the BR to ingest the tax data, create new statistical units, and update existing ones accordingly. Ideally, this is what the Census Bureau would also like to be able to do with the BLS MU data files—i.e., create a defined process for BR maintenance. While using the BLS files as an ad hoc research resource provides unquestioned value to BR in terms of quality enhancements, it is only through operational use that any sort of meaningful cost savings will be realized. Unfortunately, this kind of scenario is still quite a long way from becoming a reality. In addition to the usual culprits of budget, resources, and priorities there are some inherent issues with the data files that are going to make this difficult to achieve. This part of the paper addresses some of these challenges, specifically:

1. The completeness of the BLS MU files.
2. The timing of their availability at the Census Bureau.
3. Basic EIN-level comparison issues between the BR and the BLS files.

Each of these are described in more detail below.

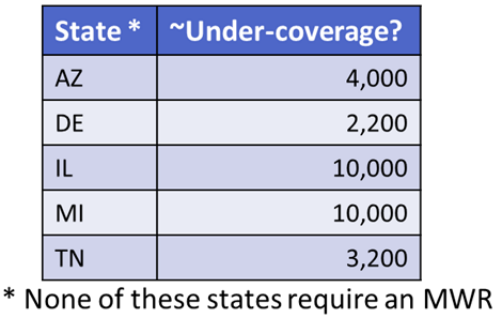
<u>Completeness of the files</u>

Some states will not allow data from their unemployment insurance (UI) system to be shared with the Census Bureau. By design then, the BLS MU files are incomplete. The table below shows the four “missing” states along with their corresponding establishment counts and employment as published in the 2012 CBP.



State	2012 County Business Patterns Data	
	Establishments	Employees
MA	171,278	3,035,897
NH	37,213	548,985
NY	527,001	7,556,521
WY	20,635	214,241
Total	756,127	11,355,644

In addition, in comparing basic establishment counts from the 2012 BLS MU data file it appears that some states may be significantly under-covered relative to the BR:



State *	~Under-coverage?
AZ	4,000
DE	2,200
IL	10,000
MI	10,000
TN	3,200

* None of these states require an MWR

It is unclear whether this presumed under-coverage is due to something definitional or to the fact that some states—including the five shown in the table above—do not require a multi-worksite report (MWR) in their UI reporting system.

In addition to the “missing” states and the potential under-coverage in others, nearly 25% of the establishments in the BLS MU file are missing either addresses, payroll and employment data values, or both.

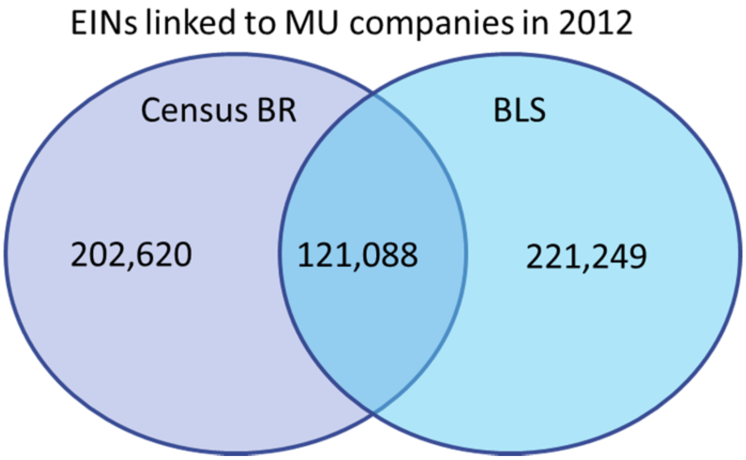
All of these file-completeness factors both limit and complicate the design of any sort of process that uses the BLS data to directly update the BR. This is particularly true for companies that are doing business in any of the “missing states”.

Timing

At the time they are made available to the Census Bureau, the BLS data files are a year behind the COS cycle. For example, the 2013 BLS files will arrive at the time that the 2014 COS is being conducted. At this point, the BR is operating on a 2014 survey year basis and the 2013 reference period is assumed to be "final" by most BR end-users. Regardless, as mentioned above in **Part 1**, despite their "lateness" the BLS files are still very useful for conducting company-level research and making decisions in the current year.

EIN matching

The EIN is the one "non-fuzzy" matching element that exists between the BLS MU data files and the BR. The presumption was that there would be a "good" comparison rate on EIN between the two data sources—i.e., that "most" MU EINs on the BLS file would also be MU EINs on the BR and vice versa[4]. Unfortunately, as the diagram below illustrates, this does not seem to be the case. Although, in terms of total MU EINs, the two data sources are relatively close—the BR has 6% fewer than BLS—there does not seem to be much overlap.



Outlined below are two different perspectives on this simple matching process that seem to underscore significant differences in the way MU companies are linked to EINs on the business lists of the two agencies.

What is a "multi-unit"?

Just prior to conducting the 2012 EC, the EINs that were on the 2011 BLS MU data file were matched to the BR. It was discovered that nearly 1.4 million (44%) of the BLS locations had EINs that were linked to SU establishments on the BR. One component of the EC questionnaires asks SU companies to report operations at multiple locations—i.e., are they splitters? Since

[4] Of course, results from *location* or establishment-level matching was expected to be another story as this would require the use of "fuzzy" elements like business names and addresses and would have to account for any definitional and special reporting unit constructs that might be present in the two business lists.

more than 2 million SUs were mailed it was determined that this universe would provide a good opportunity to validate the MU status of selected EINs found in the BLS file against the EC splitter results. So, were significantly more BLS EINs be linked to MU companies on the BR after the splitter updates had been made? The short answer is “not really”. After the EC, there were still about 1.2 million (36%) locations on the 2012 BLS MU data file that had EINs linked to SU establishments on the BR. The table below summarizes the disposition of these SUs in the 2012 EC.

Description	Count
EINs with multiple locations on the BLS file but linked to an SU on the BR	178,499
Mailed as SU in the 2012 EC	134,238
Responded in the 2012 EC	105,381
Reported operating in multiple locations *	4,787

* None of these were set up on the BR as MU due to being below size thresholds

Very few (<3%) of the cases even reported having multiple locations and none were set up as MU companies on the BR due to being below size thresholds (either 50 employees or $150,000 in annual payroll).

These results suggest that there are some inherent differences between the Census Bureau and BLS as to what constitutes a “multi-unit” company. This likely precludes the notion of “automatically” creating MU companies directly on the BR via the BLS data source and, in fact, suggests that the “Targeting Splitters” application described in **Part 1** may not bear much fruit.

Where are all of the BR EINs?

The diagram on the preceding page shows that more than 200,000 (63%) of the MU EINs that are on the BR are not even on the 2012 BLS MU file[5]. The “missing” states mentioned above only explain about 10% of them. These EINs account for nearly 450,000 establishments and 14 million employees across 100,000 distinct companies. Most of these companies are not candidates for inclusion in the COS which makes them, theoretically, ideal for updating on the BR via the BLS data. Unfortunately, the lack of a connection to the EIN in the BLS MU file raises some concerns. Where are all of these EINs in the BLS statistical unit universe? Are they SUs? (Again suggesting significant definitional differences between MU vs. SU in the statistical systems of the two agencies.) Are the 450,000 establishments in the BLS MU file under a

[5] This match was done after the 2012 EC was completed. During the EC every MU company has the opportunity to update their payroll tax EINs so the large number of mismatches is not attributable to an “out-of-date” BR. Also, in the EC, the response data for every MU company is compared to the payroll tax data at the EIN level for completeness. EIN linkage errors are identified and corrected during this process so “BR linkage error” is also not the likely explanation.

different EIN? As of yet, no serious fuzzy-matching efforts have been made at the Census Bureau to answer this last question.

Of all of the challenges outlined above, the disconnects with the EIN matching are perhaps the most troubling as they suggest that there may be some fundamental differences in statistical unit definitions between the Census Bureau and BLS. Regardless, with payroll tax data essentially being the foundation of the BR it is going to be difficult to operationalize a data source that has relatively low agreement on an EIN basis.

Summary

The BLS MU data have proven to be a valuable research and validation tool for BR analysts as they work on updating companies in the context of the COS. These data have also been instrumental in other smaller-scale, incremental improvement initiatives for the BR and the COS. However, the Census Bureau is not yet at a point where it can operationalize the BLS data sources on a large scale as a regular part of BR maintenance. The immediate goals are to develop more interactive tools around these data sources in order to facilitate access and research for BR analysts. Large-scale operational use, if feasible, is going to take more time, resources, and evaluation to achieve.