ARC

Winter 2019

Upcoming Changes to the WID Structure

Dana Placzek, Connecticut Department of Labor

The world of labor market data changes continually, and it's our job at the Analyst Resource Center (ARC) to respond to this change. With that in mind, the Structure Committee of the ARC decided on changes to the WID at three different levels.

At the level of minor changes, we expanded three description fields in WID v2.8. These changes are described in an Addendum/Errata document available at the WID Center website. One additional change was to change the numeric field length, in table *programs*, to be numeric (8,2). This allows storing longer lengths, and also fractions of time lengths.

Somewhat more significant changes will be pushed back to a v2.9 release. These changes include the addition of non-primary key fields, or additional tables that are not core. *Continued on page 8*

Training Opportunity

The Analyst Resource Center will be providing training for DBAs May 19-21, 2020 in St. Paul, MN.

This training will be provided free of charge. Participants will be responsible for travel, lodging and meals.

An official notice will be sent out to all DBAs after the first of the year when registration information will be available.

In **This** Issue

Overview of WID Structure Changes 1
Area Type Conundrum 1
National Labor Exchange.....2
Finding Data Sets.....3
In the Spotlight.....4
LMI & the Cloud.....5
Goodbye FactFinder.....5

The Area Type Conundrum

Gary Sinick, Oregon Employment Department

If you spend any time with the WID database structure, you're bound to notice that virtually every data table begins with the following three fields: stfips, areatype, and area. Together, these three fields serve to identify the area for which the data apply. 'Stfips' contains the state FIPS code. 'Area' contains the code for the specific area, and in the case of MSAs and counties, is the official FIPS code for these areas. The 'areatype' field contains a two-digit code that was created for the WID to distinguish one set of areas from another, and indicates whether the area in question is a state, MSA, county, or other type of region.

In the first version of the WID database, twenty area types were specified. The area type for MSAs was simply '02', and this worked just fine. Just fine, that is, until the results of the 2000 Census were compiled and the composition of some MSAs were changed based on new information about population, labor force, and commuting patterns.

It's easy enough to change the way an MSA's definition is described in the database. But a problem arises when some data series take longer to convert to the new MSA definitions than do others.

Continued on page 6

National Labor Exchange: A Good Source for Data

James Spector Bishop, Minnesota Department of Employment and Economic Development

With over 2.3 million active job postings, National Labor Exchange (NLX) has a lot of potential as a data source. However, to make use of it, it is important to have a bit of background on it, and an understanding of its advantages and pitfalls. The NLX is an online job posting network run by Direct Employers, connecting employers, state job banks, and workers. It functions as a huge database of active jobs openings which can be accessed through their API. The NLX collects these jobs from a variety of sources, including members of Direct Employers, as well as daily uploads from state job banks.

As a source of data, the NLX has a lot of things going for it. As mentioned above, it is a large dataset, allowing for statistically significant samples. It contains detailed occupational information in the form of multiple O*NET codes for most job postings. Postings on the NLX are currently live, with inactive postings being constantly removed. This means that it can function as a snapshot of job vacancies at any given moment, providing the potential for real time analysis. The fact that it is constantly being updated also means that it could be used to identify short term labor market trends as well as mass hiring events. Perhaps the most useful qualities of the NLX are its granularity and versatility. By granularity, I mean that it contains individual job postings, so it is open to being aggregated however we want. By versatility, I mean that so long as we have the data, we can always ask it new questions and analyze it in new ways. These qualities benefit from the fact that the NLX is public data, consisting of things that employers themselves posted for all to see. It is not confidential nor proprietary and thus its usage is not restricted in the same ways that labor market surveys, with their black boxes of confidential respondent information, are.

For these reasons, the NLX can be used, and is indeed already being used, to mine job postings for data. CareerOneStop currently uses the NLX to generate its list of in-demand professional certifications. It does this by analyzing hundreds of thousands of NLX job postings to identify mentions of different certifications in order to gauge their relative popularity among employers. Though we are still working on developing other applications for the NLX data, we have created several tools capable of analyzing and categorizing NLX jobs based on dozens of characteristics including: the pay and benefits offered, the level of education and experience they require, veteran's preference, and more! This indicates that there is a lot of information that can be gathered from the NLX. We hope to take advantage of the NLX in the future to build useful tools and data to share with other ARC members.

However, though it has many positive qualities, the NLX also has a number of limitations as a data source. Of greatest concern is the fact that it is not very representative of the job market. It underrepresents some locations, such as Louisiana, and over represents others, such as Minnesota. Similar problems exist among employers. The NLX mostly lacks jobs from Walmart, the largest private employer in the country, while having far more jobs from Oracle than is possible. In addition, based on a comparison to the Minnesota Job Vacancy Survey (a semiannual survey of job openings in Minnesota), the NLX is not representative in terms of occupations, skewing more towards high pay, high education. However, this comparison also found that some occupations, such as trucking, are very well represented in the NLX, indicating that there may be subsets of the NLX that provide a fairly complete and accurate picture.

The NLX also has limitations in terms of its quality and usability. The live nature of its API means that it cannot be used to gather historical data about job postings over time. It also only provides a few fields of data for each job, forcing us to rely of job descriptions to parse out job characteristics. Unfortunately, the job descriptions themselves can lack detail, often not including pay or educational requirements. Additionally, job descriptions must be gathered from the API one at a time, making data collection very time consuming.

Continued on page 7

Finding Datasets

Amanda Rohrer, Minnesota Department of Employment and Economic Development

There's so much information on the internet that one of the big challenges is just finding the right sorts of results and making sense of them without having to have a person read the content. To combat this problem, there are structures in place for assigning machine-readable metadata. In essence, that's data invisible to a person browsing the site, but which lays out key features of the types of information in the page or accessible from the page. That allows websites to use automation to find and organize similar content without requiring that the presentation to the public use the right search terms or be structured in any particular way. This is commonly used for some types of content – news articles and recipes both rely on it. For articles, it allows them to be presented seamlessly in different publications or aggregators, and for recipes it allows apps to save and store ingredients, instructions and nutritional information neatly.

The available structures have options for all kinds of content like people and organizations and events. More and more datasets are being described and organized using this kind of metadata, so it may become more of a priority for state agencies. From a technical standpoint it's not difficult – templates and tutorials exist and some content management software (Socrata, WordPress plugins) will help content creators fill in and manage this information without ever having to alter code.

Why does it matter?

Indexing information with machine-readable metadata requires extra work – even if there's a product that's generating the code automatically, knowing that it's necessary and building that in to the site and ensuring that it's properly implemented and maintained is another layer of web design and maintenance that may not be high on a list of priorities for small organizations. And once that index is created, there still have to be users who are accessing the information through the metadata to make that extra effort worth the time. Right now, the most obvious way you'd see metadata benefit your site is in search engine results. All the major search engines will use it to improve the blurb that appears in their results. While it likely doesn't affect search rankings, having a more relevant description of your content may make it more likely that searchers will click on the link.

Another way it may become increasingly relevant is with the rise of services such as Alexa – when users are searching by voice, the information needs to be organized such that only a couple of results with high confidence are returned and the bot can go straight to the important information.

Finally, and most relevant to LMI offices, there are more dataset aggregators being developed or expanding that rely on this type of structure. The two major ones are Google's Dataset Search and the federal Data.gov.

Open Data

In 2013, Executive Order *Making Open and Machine Readable the New Default for Government Information* set standards for federal agencies and how they manage data. The subsequent memo defined those requirements with more clarity. While the 2019 Evidence Act mandated further changes, since 2013 it has been mandatory for federal agencies to "use machine-readable and open formats for information as it is collected or created".

The practical implication for this is that the federal government has very specific requirements defined for how to apply metadata and their own platform for sharing data (Data.gov) requires and relies heavily on that metadata for their site and the data harvesting tools they build off it. Data.gov allows local governments to include their data in the listing, but they have to structure their data according to similar standards as the federal agencies.

Continued page 9

In the Spotlight

Mike Peery works for the Montana Department of Labor & Industry, and is a member of the Analyst Resource Center Consortium.

How long have you been involved in the world of LMI? I'm approaching almost 29 years in the world of LMI. I have worked in some capacity in all the BLS Cooperative program. . . including the Survey of Occupational Injuries & Illnesses (SOII) and the Census of Fatal Occupational Injuries (CFOI). I've also been responsible for LMI Outreach in Montana to local Job Service Offices, businesses and schools.

Are you originally from Montana? I was born and raised in Montana. I took a job as a retail manager after college and was transferred to Oregon. I lived in multiple cities in Oregon in the late 80's, but ultimately moved back to Big Sky Country.

What is your role on the ARC? I currently serve on the Education and Training Committee with the ARC. This has been a great fit for me as I served on the LMI Institute Education & Training Committee years ago and have been involved in several LMI Outreach projects in Montana.

That is your current job title? I am currently the Labor Market Information Director in Montana.

What is the most rewarding aspect of your job? The most rewarding aspect of my job is knowing that at the end of the day, we are helping move the dial to improve the lives and workplaces of the citizens and businesses in Montana. Whether it's helping someone find a job or retrain, or give a student a pathway to a career, or assist a business with recruitment, retention, and on the job training…we are impacting lives in a positive manner each and every day. And all of our successes are tied back to our high quality LMI and Career data. We've got them, and we're not afraid to use them.

What is the most frustrating or challenging aspect of your job? The most frustrating aspect of my job, as I'm sure it is for many of my colleagues is funding. We perform essential functions in career and workforce development, and we possess the most crucial element to success in these endeavors in the form of data. Yet somehow, the funding to provide and assist in "data driven decision making" is grossly underfunded.

What is the strangest/interesting job you have ever had? Strangest job I had was picking up and driving a van full of sober river floaters and their raft to a launch site, driving down river several miles, and then waiting for the group of river floaters (now extremely intoxicated). I loaded them and the raft up and drove them back to their homes. In hindsight, this was a responsible way for them to recreate and get home safely. I was paid $50 for each excursion…which was big money to me back in the 80's.

What about your family? I have a wife and 3 grown sons from ranging in age from 27 to 30 years old. Sadly, they are all single and I have no grandchildren yet. This is a frequent topic of conversation with them.

LMI and the cloud – FedRAMP

Matt Steadman, Utah Department of Technology Services & LEWIS Lead Developer

In the latest cooperative agreement from BLS there is a provision for using cloud services with confidential data. Section T:17 of the Administrative requirements reads:

"State use of a Cloud Service Provider (CSP) to service BLS confidential information must be through an authorized FedRAMP vendor and the vendor's FedRAMP package must be reviewed and approved by the BLS prior to use. Confidential information must have defined access controls and be encrypted at rest and in transit to prevent unauthorized access. Only FIPS-validated cryptography is approved for use in encrypting Federal information. Any employee of a CSP who will require access to confidential information in an unencrypted environment (e.g., to provide support, aid in migration, troubleshooting) must be a designated BLS agent and complete required training."

This is an exciting opportunity for people to take advantage of new technologies that have not been approved in the past. However, this description may be rather confusing and raise more questions than answers for people. This article is a dive into what the federal guidelines are when it comes to using these services for confidential data, and how BLS is interpreting those guidelines.

The goal of the Federal Risk and Authorization Management Program (FedRAMP) is a government-wide "do once, use many times" framework that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services, that saves cost, time, and staff required to conduct redundant Agency security assessments. The FedRAMP marketplace, on fedramp.gov, provides a list of products that have met these requirements. BLS asks that the associated BLS Regional Office be notified of the CSP name, the FedRAMP Package ID, and the intended use of the product within the state for approval.

Once approved, states will need to ensure that the product is used in a properly secured manner. All data should be encrypted, using at minimum encryption that meets the FIPS 140-2 standard, both in transit and at rest and only individuals who have signed BLS Agent Agreements and completed BLS confidentiality training should be allowed access to unencrypt the data. While this includes employees of the CSP, not just state employees, it is possible using public-key encryption to secure the data so that not even the CSP will have access to it. Should the need arise to give access to a CSP staff member, for troubleshooting or other purposes, they would need to complete the BLS agent agreement and training.

Goodbye FactFinder, Hello data.census.gov

Amanda Rohrer, Minnesota Department of Employment and Economic Development

The Census Bureau has settled on a redesign for their data distribution products. While the underlying structure of the Census data (the tables and table identifying numbers) won't change, the presentation of them to the public will. The intent is to simplify the use of Census data, but LMI offices tend to be specialty users – they have expertise in the data and very specific needs and often use the same data year after year. Changing the download formats will have an impact on LMI offices – not only will state employees have to learn the new system, there may be structural changes to the access or output that will require processes to be reworked. The Census Bureau has a Release Notes document that is maintained with the most current information and describes the features and status of changes in detail. Read the entire article here.

The Area Type Conundrum - *continued*

Moreover, historical data that is revised to conform to the new MSA definitions may not stretch all the way back to the beginning of the series. So, if a state wishes to keep data that uses the older MSA definitions, they will have what amounts to an undocumented series break (at least from the standpoint of the database).

What to do? Areatype to the rescue! The WID structure was updated so that MSAs based on the 2000 Census definitions were now identified with a code of '21', while '02' could still be used for MSAs using the 1990 Census definitions. At the same time code '21' was added, areatype codes were also added for 2000 definitions of Micropolitan Statistical Areas (22), Metropolitan Division (23), Combined Statistical Area (24), and four additional codes related to NECTAs (New England City and Town Areas).

Fast forward ten more years. The 2010 Census has been conducted, and new MSA definitions are published. What to do now? Add more areatypes! MSAs based on the newest definitions are now coded as '31', and four more 2010-related codes were also added.

So, what started out as a very straightforward situation (the MSA areatype is '02'), became a bit more complicated (MSAs could be areatype '02', '21', or '31'). And the situation is set to repeat itself when the census is conducted again next year. To make things even more interesting, the OMB has declared that MSA definitions will be updated every *five* years instead of every ten, which would potentially double the rate at which new MSA definitions need to be accommodated in the WID. (NOTE: According to the Bureau of Labor Statistics, data series published for MSAs will only use updated definitions from the decennial census, so that particular bullet appears to have been dodged, at least for now.)

As mentioned above, the areatype field is two digits. At the time this field was defined, twenty codes had been identified, and it probably seemed very unlikely that more than 100 would ever be needed. However, after reserving an additional twenty codes for state-specific use (50-70), and adding a big block of codes after each decennial census, it's clear that, under the current structure, areatype will slowly but surely run out of codes.

The proliferation of MSA areatypes has created another cause for concern. When the areatypes change, states need to update their WID databases with the new values. But in many cases, the new MSA definitions are the same as the old definitions. Or, only data using the new definitions is populated in the WID, and so only one MSA areatype is needed. If the areatype field is only being used behind the scenes to look up the names and descriptions of the MSAs, it's easy to imagine that some states might put a relatively low priority on updating the areatypes, as it may not have any practical effect on how data is displayed or used within applications.

So, while many states are likely using the current code of '31' as the areatype for their MSAs, others may not have switched from '21', and some may even still be using '02'. This lack of consistency may not cause any problems within each state, but when states are sharing data, trying to use common applications, or making their data available through an API, this situation has the potential to be a very confusing mess.

Part of the problem may be a limitation of the WID structure itself. The current structure that defines each area by the combination of its state FIPS, area type, and area dictates that if we want to be able to distinguish MSAs defined using the 2000 Census from MSAs defined using the 2010 Census, we don't have any options but to add more areatypes. But is MSA (2000) really a different type of area from MSA (2010)? Or is what we really need a way to distinguish one *version* of an areatype from another? If an 'areatype version' field were added to the structure it would solve this problem more elegantly. This structure would go back to using a single areatype code for MSAs (perhaps even the original '02'), then an additional field would indicate which version of the MSA definition is being used (1990, 2000, 2010, etc.). This approach would have the added benefit of being able to easily select all MSA data that are using any version (past or present) of the MSA definitions.

Continued on page 7

The Area Type Conundrum – *continued*

Introducing such a fundamental change to the way that areas are defined in the WID, however, would be a significant effort to implement, as it would likely require modifications to any applications that are built using the WID database. An update such as this would be considered a "breaking change", and would only be rolled out as part of a major update to the WID structure (i.e. version 3.0). Whatever its merits might be, this change to the structure is not something that could be made quickly, but should be considered as a potential part of a package of significant updates that will be included in the next major version. In the meantime, it would behoove states to have a look at the areatype codes currently being used in their WID databases to make sure that they conform to the current 2.8 version of the WID.

National Labor Exchange: A Good Source for Data - *continued*

On the bright side, most jobs come with multiple occupation codes and much of the missing information can often be extrapolated based on these. Finally, jobs on the NLX database represent job postings rather than job openings, with a single posting potentially representing multiple open positions, and some postings for high turnover jobs remaining active for years on end.

In conclusion, though the NLX has its limitations, it has a lot of advantages as well, and its breadth and timeliness make it hard to ignore as a potential data source.

Comparison of the NLX to the Labor Market

Sample	National Labor Exchange	MN Job Vacancy Survey
Median Wage	$16.50 - $24.60 per hour	$15.01 per hour
Percent of jobs requiring a college degree (associates, bachelors, or advanced)	30.4% - 40.6%	23.5%
Percent of jobs requiring a license or certification	36.4%	33.8%
Largest Occupations in the Sample	Truck Drivers (533032) Registered Nurses (291141) First-Line Supervisors of Retail Sales Workers (411011)	Retail Salesperson (412031) Food Prep and Service (353021) Personal Care Aides (399021)
Most Overrepresented Major SOC Group in the Sample When Compared to the other Sample	Computer and Mathematical Occupations (150000)	Food Preparation and Serving Related Occupations (350000)
Largest Employer	Find A Trucker Job.com (fatj.com)	*Unknown*

Overview of Upcoming Changes to the WID Structure - *continued*

One such change is a request for an occupational "level" field, to aid in rolling occupations up. This will be added to all of the occupational lookup tables. Another request for a way to identify high-demand occupations led us to revise the table *iospecialid*. Originally conceived to identify green jobs, this table can now be used to track STEM jobs, high-demand jobs, and any other type a user can think of. These tables, along with prototypes of new *qwi* and *esapplicant* tables, will be placed on the Non-standard pages of the WID Center. Once they've been tested and found to work, they will be added to a WID v2.9 release. This will take a minimum of a year, so it will be at least that long before we release v2.9

The big news to come out of our October 2019 meeting is the decision to draft the next major release: WID 3.0. All of the changes discussed above are minor changes that should not have an effect on existing applications. By contrast, a major release will affect any and all applications written against the WID; these will need to be re-written. Due to the scope of these changes, we will give State DBAs and other stakeholders plenty of chances to comment on the changes made.

The main motivation for a major release is the realization that our method of handling geography is inadequate, and we are running out of area type codes. A version 3 also means that we are no longer constrained by naming conventions and other standards from the 1990s; this is our opportunity to create consistent naming and structure conventions. We can also develop the WID in conjunction with our push towards standard APIs, wherever this makes sense. Other issues have also cropped up over the years, only to be pushed off "when we publish v3.0".

However, the process of creating and approving a major release will be slow and deliberate. First, the Structure Committee has to come up with a draft database structure that it likes, and then present this to the larger ARC group for discussion and revisions. Next, the draft will be released to ETA and BLS, and other stakeholders, for their comments and recommendations. This will include input from State LMI directors and DBAs. This entire process will likely take two to three years.

While such a major change may seem scary, we are hoping that, with enough advance warning and participation from all interested parties, we can create a new Workforce Information Database that will serve the LMI community for as long into the future as version 2.x has in the past.

Finding Datasets - *continued*

Schema.org

The major search engines (Google, Yahoo, Bing, and Yandex!) collaborated to start Schema.org which lays out a structure for web content generally. Their products use this information on websites to improve how meaningful the site information is in search results. There are similarities with the federal standards for data but the Open Data Project provides mappings to other structures, including Schema.org.

Google Dataset Search

Like many Google products, Dataset Search is in Beta. If you search for a topic or an organization it will come up with a list of sources and describe them with consistent fields – when it was last updated, who the data owner is, the format of the content, the time period, and a description. Searching for CareerOneStop will give a list of data products and related organizations like the St. Louis Fed are heavily represented, but few state agencies currently show up in this tool.

Implications

Being an established federal standard and having products relying on machine-readable metadata means that structuring content behind the scenes is getting more common and more necessary and may eventually be a requirement for states. Having more of it done automatically with the other products we're using (content management systems or other) means it is possible agencies are describing information without a full awareness of where metadata is being picked up and used and may never review how their data is being presented at sites beyond their control. Understanding the parts of websites that are not visible and knowing what the larger organization's strategy is for producing and managing metadata may allow LMI offices to establish long-term goals that can be implemented incrementally instead of reacting after technology or regulations change.

ARC Consortium Meeting October 2019

James Spector-Bishop, Nicole Kennedy, Steve Hine, Al Sylvestre, Andy Condon, Amanda Rohrer, Dana Placzek, Matt Steadman, Steve Duthie, Mike Sylvester, Gary Sincick, Bill McMahon, Mike Peery, John Pearce, Barb Ledvina, Christopher Robison, Joe Jaehnke, and Kevin, Doyle.

ARC Newsletter

Editor: Barbara Ledvina

Thank you to Amanda Rohrer, Minnesota; Dana Placzek, Connecticut, Matt Steadman, Utah, Gary Sincik, Oregon, James Spector-Bishop, Minnesota and Mike Peery, Montana for their contributions to this edition.

The Workforce Information Database is a normalized, relational database structure developed for the storage and maintenance of labor market, economic, demographic and occupational information. The Analyst Resource Center is responsible for the structure development, update, and maintenance of the Workforce Information Database. Current members include representatives: Minnesota (lead), Connecticut, Florida, Iowa, Michigan, Montana, Nevada, North Carolina, Oregon, Utah, Virginia and Wisconsin.