GEORGETOWN UNIVERSITY
Center
on Education
and the Workforce
McCourt School of Public Policy

To: Byron Auguste, deputy assistant to the President and deputy director of the White House National Economic Council

From: Anthony P. Carnevale, director of the Georgetown University Center on Education and the Workforce

Date: November 10, 2014

Subject: **The Need for Public Investments in an Open Source Real-Time Data Repository**

We are in the midst of massive innovations in data systems that promises to finally link information from postsecondary and training institutions with labor market demands. So far, there are three major connections between the educational world and the labor force. The first is the Statewide Longitudinal Data Systems (SLDS), which links unemployment insurance wage records[1] from all 50 states with transcript data from education institutions. The second has been developed by private vendors: Internet job postings data (a.k.a. real-time data), which provide the most current snapshot of American labor markets available. The third, which is still developing, is made up of a variety of sources that use data from resumes stored at public and private institutions.

Yet despite these rapid and critical innovations, the federal government has largely remained on the sidelines in developing, validating, and integrating this wealth of information. Internet job postings have been a notable missed opportunity for public and private collaborations. In the '90s, private firms began submitting their job openings to the U.S. Department of Labor to comply with new federal workplace discrimination regulations. The Department of Labor combined these daily feeds with job listings from public employment services to create America's Job Bank. That project was discontinued during the Bush Administration. As a result, what began as an organic public-private collaboration—made up in part by public records—has evolved into a proprietary resource for private businesses.

Private for-profit entities have performed a valuable service by developing job openings data, but a public role is sorely needed. What's more, the value of real-time job openings data will remain limited until job openings data is connected to data from publicly administered and/or publicly funded institutions, information currently protected by privacy regulations.

The federal government should create an open-source platform that would integrate these two data sources. This would not only serve the public's interest in facilitating research and informing economic

[1] One key failing of the emerging data system linking education and careers is the absence of an occupation al designation from the Standard Occupational Classification code in the wage record data. An individual's occupation is much more tightly tied to their education field of study than their industry. With the exception of Alaska and soon Louisiana, wage records are incapable of connecting field of study in education programs to occupation. Adding occupation to wage records could be done with a minimal increase in compliance burdens for employers. Employers already provide occupational information in their EEO and Disability data submissions. Adding occupation to wage records could partially reduce EEO and Disability compliance burdens.

policy decisions, it would also benefit the private companies which collect the data. As of now, each private vendor collects its own job postings information, a duplication of effort and a major source of noise, misunderstandings, and inaccuracies.

There are other basic flaws in the job openings data[2] that can only be overcome by such an open-source approach. For example:

- There are natural limits to the value of job postings data used all by themselves. Individual job postings rarely include more than 50 words for parsing, only about 10 of which actually impart useful information.
- The aggregation of job openings varies widely. For example, Burning Glass, a leading private vendor of job openings data, includes roughly 1.7 million advertisements. Other vendors, such as Job Openings and Labor Turnover Survey (JOLTS), list roughly 3.3 million; another vendor, Help Wanted Online (HWOL), lists well over 4 million openings. These numbers suggest considerable duplication.
- This duplication occurs, in part, because some job boards are merely aggregators of other job board data, and de-duplication software varies in its effectiveness.
- Occupation data is reliable for more than 80 percent of real-time job ads, but only 40 percent of those ads mention a specific education requirement.
- Coverage is biased by sector, with industries such as agriculture, construction , utilities and manufacturing tending to be under-represented.
- Industry definitions remain incomplete, and overall education requirements are biased upward: positions requiring a Bachelor's or a graduate degree are over-represented, while fewer than half of the postings are for positions requiring less than a Bachelor's degree.

There are lots of models for public/private collaboration in data development and use. One exemplary example is the Employment and Training Administration's Occupational Information Network (O*NET), which has built a base data set and tools like "My Next Move" without discouraging vendor use and open market competition. Such examples show that the federal government should be able to maintain a database, create primary tools, and establish an open source platform while still leaving vendors free to compete with each other in developing short-term projections and highly enhanced data mining. An open source public presence would also lower barriers to entry into the data development market for both private vendors and public groups.

Because they command privacy-protected administrative data and access to clients, especially clients in need, public agencies are uniquely positioned to integrate education and labor market data because they touch individuals at crucial decision-making junctures in their lives—for example, when they apply for Pell Grants or college loans, when they transition from school to work, when they apply for unemployment benefits, when they muster out of the military and when they come into contact with the criminal justice system. New data streams like real-time data or resume data will be most powerful as complements rather than substitutes for the much more powerful privacy protected administrative data flows through public agencies. Transcript data and wage records are obvious current examples in the education and training policy context.

---

[2] See our recent report "Understanding Online Job Ads Data: A Technical Report," at https://georgetown.app.box.com/s/nre5ybcw97e8gpyq502w.

On the other hand, an exclusively for-profit development strategy generally minimizes public investments at the public core of a robust information infrastructure. Without public investment, most federal, state and local agencies are likely to be very slow as developers and informed consumers of new data. Colleges, career centers and other employment service providers, community-based organizations, disability organizations, and the criminal justice system have already fallen way behind in using this information to match education and labor market services with good jobs.

The public role is *necessary* to promote educational equity and economic mobility. Information and tools developed by private, for-profit firms will ultimately be accessible only to those who are able to pay for it. For-profit institutions will naturally follow the money, wherever it leads. The ongoing dust-up with for-profit colleges through the Gainful Employment regulations perfectly illustrates the challenges created when the public sector leaves the field to in deference to private for-profit entities.

Absent a public role in the development of these new data feeds, disadvantaged individuals and the agencies that serve them will be last in line as developers and knowledgeable consumers of new capabilities like real-time data. Disadvantaged groups, such as immigrants and veterans, already rely on distorted information trickling down to them through their peers and popular culture. Information tools and counseling can help these strivers overcome these barriers, and would be essential supports as they attempt to climb social and economic ladders.

In my view, the Census Bureau's Local Employment Dynamics (LED) and Longitudinal Employer Household Dynamics (LEHD) staff would be best suited to integrate real-time data into an open source public platform. LEHD/LED has already established a credible public reputation as an innovative integrator of public data sources. Using Census data, LEHD/LED has imputed educational requirements to jobs and wages, a particular weakness of the private real-time data sets. LEHD/LED appears to have the best team available to link occupation and education data, a critical missing piece in wage record data. In addition, LEHD/LED's connection to commuter patterns would make it an ideal tool for integrating education and economic development data.

In summary, the federal government can create great value by collecting, de-duplicating, and authenticating the data field reliability of these data.[3] Using the O*NET model, this activity would provide an open-source feed of consistent quality in a world where there appears to be no competition on aggregation activities. In doing so, the federal government would lower cost to states that have developed data tools on their own, while taking an important step to improve data quality.

Anthony Carnevale
Research Professor and Director
The Georgetown University Center on Education and the Workforce
3300 Whitehaven St. NW, Suite 5000, Washington, DC 20007
tel: 202.687.4971
fax: 202.687.3110
cew.georgetown.edu

[3] Duplication occurs partly because some job boards are merely aggregators of other job board data.