Better Together: Three Models for Combining Public and Private Data to Create New Research Data Sets

Eddie Tzu-Yun Lin

University of Chicago

Follow this and additional works at: https://repository.upenn.edu/admindata_conferences_presentations_2017


This paper is posted at ScholarlyCommons. https://repository.upenn.edu/admindata_conferences_presentations_2017/9
For more information, please contact repository@pobox.upenn.edu.
Better Together: Three Models for Combining Public and Private Data to Create New Research Data Sets

This presentation is available at ScholarlyCommons: https://repository.upenn.edu/admindata_conferences_presentations_2017/9
Better Together: Three Models for Combining Public and Private Data to Create New Research Data Sets

Eddie Tzu-Yun Lin
tlin2@uchicago.edu
Outline

- WDI Overview
- Challenges of Labor Market Data
- Open-Skills Project
- Combining Private and Public Sector Data
  - Use Public Sector Data for Validating Private Sector Data
  - Use Public Sector Data to Extract New Data from Private Sector Data
  - Use Private Sector Data for Creating New Public Sector Data
- What can you do today?
The Workforce Data Initiative
What We’re Doing

- Open Data on Skills and Jobs
- Opening Training Outcomes Data
- Supporting New Developer Resources and Third Party Tools
Challenges of Labor Market Data

- Lack of geographic and temporal granularity of survey-based labor market data sources
- Lack of intra-occupational specificity of national labor market taxonomies/ontologies
- High frictions to data access for more granular administrative data sources
- Lack of technical capacity among social science researchers for working with large, unstructured data sources
- Lack of transparency in methods for generated datasets from most Real-time Labor Market Information data vendors
Open Skills Cooperative
- Pooling public and private data into a private data store
- 9 months of negotiating unique legal and governance structure

Open Skills Research and Development Group
- Data scientists, machine learning and AI researchers
- Economists
- Taxonomists, and other subject matter experts

Open Skills API
- Open aggregate data resource of raw and normalized skills & jobs, and how they relate to each other in time and place

Open Skills Developer Partners
- Building applications on top of the Open Skills API
- Supporting veterans’ spouses, career one-stop-shops, & industry use cases

Open-Skills Project

Data@Work Research Hub

MACHINE LEARNING

Open Skills Database

O*NET

Data@Work Research Hub

Open Skills Developer Partners
Three Models for Using Private and Public Sector Data

Public Sector
- BLS: JOLTS, OES
- O*NET
- Census Bureau

Private Sector
- Data@Work Research Hub
- Open-Skill API
- EMSI Data
# Data@Work Research Hub Dataset

<table>
<thead>
<tr>
<th>CBSA Name</th>
<th>State Code</th>
<th>Title</th>
<th>SOC Code</th>
<th>Onet Skills 1</th>
<th>Onet Skills 2</th>
<th>Onet Skills 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detroit-Warren-Dearborn</td>
<td>MI</td>
<td>automotive production supervisor location relocat...</td>
<td>43-5053.00</td>
<td>1 rules</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>Hartford-West Hartford</td>
<td>CT</td>
<td>dispatcher public safety</td>
<td>33-3051.01</td>
<td>1 transportation</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>Minneapolis-St. Paul</td>
<td>MN</td>
<td>digital strategy director</td>
<td>11-2031.00</td>
<td>2 design</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>Philadelphia-Camden</td>
<td>PA</td>
<td>database coordinator sql server</td>
<td>15-1141.00</td>
<td>1 programming</td>
<td>science</td>
<td>design</td>
</tr>
<tr>
<td>Auburn-Opelika, Al</td>
<td>AL</td>
<td>customer service representative</td>
<td>43-5071.00</td>
<td>2 guides</td>
<td>self</td>
<td>microsoft office</td>
</tr>
<tr>
<td>Bastrop, LA Micro Area</td>
<td>LA</td>
<td>pro</td>
<td>49-3023.02</td>
<td>2 transportation</td>
<td>auto</td>
<td>null</td>
</tr>
<tr>
<td>San Jose-Sunnyvale</td>
<td>CA</td>
<td>insurance agent opportunity</td>
<td>41-3021.00</td>
<td>1 auto</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>Sacramento-Roseville</td>
<td>CA</td>
<td>multiple positions available entry level manageme...</td>
<td>11-2022.00</td>
<td>5 route</td>
<td>player</td>
<td>sales and marketing</td>
</tr>
<tr>
<td>Chicago-Naperville-East</td>
<td>IL</td>
<td>global business communications specialist</td>
<td>11-2021.00</td>
<td>1 linkedin</td>
<td>templates</td>
<td>facebook</td>
</tr>
<tr>
<td>Trenton, NJ Metro Ar</td>
<td>NJ</td>
<td>administrator customer support clerk</td>
<td>41-2031.00</td>
<td>1 clerical</td>
<td>levels</td>
<td>null</td>
</tr>
</tbody>
</table>
Use Public Sector Data for Validating Private Sector Data
Representativeness Analysis

Single Year Representativeness Summary Data Overview (2013)

<table>
<thead>
<tr>
<th></th>
<th>OES</th>
<th>JOLTS</th>
<th>Data@Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Employment</td>
<td>Job opening</td>
<td>Job posting</td>
</tr>
<tr>
<td>Counts</td>
<td>169M</td>
<td>90M</td>
<td>2.4M</td>
</tr>
<tr>
<td>Geographic Granularities</td>
<td>586 MSAs</td>
<td>4 Regions</td>
<td>1004 CBSAs</td>
</tr>
<tr>
<td>Major Groups</td>
<td>22 (ONET)</td>
<td>26 (NAICS)</td>
<td>23 (ONET)</td>
</tr>
</tbody>
</table>
Job Distribution by State: Data@Work vs. OES

Correlation: 0.8738

NOTE: Based on 2013 data. Does not include AK and HI.
Percent Change in Data@Work Job Postings (contributed by CareerBuilder) vs. JOLTS Job Openings

NOTE: Based on 2011-2014 data.
Use Public Sector Data to Extract New Data from Private Sector Data

- Public Sector Data
  - Occupation Normalization
  - Skill Normalization
  - Occupation Classifier
  - Skill Classifier
- Private Sector Data
  - Open-Skill API
  - Data@Work Research Hub
Use Private Sector Data for Creating New Public Data

Private Sector → DSaPP → Public Sector → Open Data

Resume Data → Job Title Normalizer Occupation Classifier → RSI Form
What can you do today?

- **Sign up** to be part of the Data@Work researcher community on Slack
- **Explore the data** on the Data@Work Research Hub
- **Ask questions and provide suggestions** for additional datasets to the Data@Work on Slack
Occupation Distribution Correlation by State

- Mean: 0.549901
- Standard Deviation: 0.054917
- Minimum: 0.446043
- 25%: 0.514191
- 50%: 0.556185
- 75%: 0.585207
- Maximum: 0.712504

NOTE: Based on 2013 data. Does not include AK and HI.
Use Public Sector Data to Extract New Data from Private Sector Data

Open Skills API
A complete and standard data store for canonical and emerging skills, knowledge, abilities, tools, technologies, and how they relate to jobs.

Created by Work Data Initiative
See more at http://www.dataatwork.org

<table>
<thead>
<tr>
<th>default</th>
<th>GET</th>
<th>/jobs</th>
<th>Job Titles and Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GET</td>
<td>/jobs/{id}</td>
<td>Job Title and Description</td>
</tr>
<tr>
<td></td>
<td>GET</td>
<td>/jobs/{id}/related_skills</td>
<td>Skills Associated with a job</td>
</tr>
<tr>
<td></td>
<td>GET</td>
<td>/jobs/{id}/related_jobs</td>
<td>Jobs Associated with a job</td>
</tr>
<tr>
<td></td>
<td>GET</td>
<td>/jobs/autocomplete</td>
<td>Job Title Autocomplete</td>
</tr>
<tr>
<td></td>
<td>GET</td>
<td>/jobs/normalize</td>
<td>Job Title Normalization</td>
</tr>
<tr>
<td></td>
<td>GET</td>
<td>/jobs/unusual_titles</td>
<td>Unusual Job Titles</td>
</tr>
<tr>
<td></td>
<td>GET</td>
<td>/skills</td>
<td>Skill Names and Descriptions</td>
</tr>
<tr>
<td></td>
<td>GET</td>
<td>/skills/{id}</td>
<td>Skill Name and Description</td>
</tr>
<tr>
<td></td>
<td>GET</td>
<td>/skills/{id}/related_jobs</td>
<td>Jobs Associated with a Skill</td>
</tr>
<tr>
<td></td>
<td>GET</td>
<td>/skills/{id}/related_skills</td>
<td>Skills Associated with a Skill</td>
</tr>
</tbody>
</table>