Explorations in Data Innovations—Can Machine Learning Support Data Catalog Development?

SUMMARY

In 2021, the Chief Evaluation Office (CEO) funded contractor Westat Insight and their partner American Institutes for Research to conduct the Explorations in Data Innovations project. This use case study was designed to explore options around employing machine learning to create and maintain a public-facing, labor-related data catalog. Data catalogs are one way to address a goal of CEO to make data sources and initiatives in the employment and training field more accessible to researchers. A publicly available data catalog would help labor researchers more easily understand the full range of available labor-related datasets that could potentially answer pressing questions on labor-related policies and programs. The development and maintenance of data catalogs is labor-intensive; this project's goal was to determine the feasibility of different automation options for building data catalogs. In this effort, the study team explored relevant literature, piloted a manual data catalog assembly process, developed options for each step of the data catalog process, and consulted with a technical working group of computer science experts.

This Department of Labor-funded study contributes to the labor evidence-base to inform data, methods, and tools that build evidence on Departmental programs and policies and addresses Departmental strategic goals and priorities.

RESEARCH QUESTIONS

- To what extent can automated machine-learning algorithms be used to identify new data sources and initiatives on employment and training outcomes?

KEY TAKEAWAYS

- Insights from the pilot study and feedback from technical working group members suggest that DOL may not be able to automate the data catalog process at this time.
- Data sources have diverse structures and metadata available, making the development of an automation program difficult, and employing automations to even portions of the data catalog development process would require a large investment in staff and computing resources.
- Artificial intelligence is a rapidly evolving field. Literature suggests there may be many opportunities to support automated data collection in the future, including the use of generative artificial intelligence (Cherradi et al., 2023; Yarlagadda, 2017).
- When using machine learning to produce public-facing products, federal agencies may need to use a mix of staff with different skill sets, including data scientists, website developers, experts in cloud computing, and subject matter experts.

SEE FULL STUDY

Explorations in Data Innovations—Can Machine Learning Support Data Catalog Development?

TIMEFRAME: 2021-2024
SUBMITTED BY: Westat Insight
DATE PREPARED: June 2024

SPONSOR: Chief Evaluation Office
CEO CONTACT: ChiefEvaluationOffice@dol.gov

The Department of Labor's (DOL) Chief Evaluation Office (CEO) sponsors independent evaluations and research, primarily conducted by external, third-party contractors in accordance with the Department of Labor Evaluation Policy. CEO's research development process includes extensive technical review at the design, data collection and analysis stage, including: external contractor review and OMB review and approval of data collection methods and instruments per the Paperwork Reduction Act (PRA), Institutional Review Board (IRB) review to ensure studies adhere to the highest ethical standards, review by academic peers (e.g., Technical Working Groups), and inputs from relevant DOL agency and program officials and CEO technical staff. Final reports undergo an additional independent expert technical review and a review for Section 508 compliance prior to publication. The resulting reports represent findings from this independent research and do not represent DOL positions or policies.