International Journal of Population Data Science (2018) 3:3:156

International Journal of Population Data Science

Journal Website: www.ijpds.org

Ligo: An Open Source Application for the Management and Execution of Administrative Data Linkage

Lawrance, G[1], Hernandez, RP[2], Mamakani, K[2], Khan, S[2], Hills, B[3], Yip, H[3], and Marrville, C[2]

[1]DataBC, Integrated Data Division, Jobs, Trade & Technology, Government of BC

[2]Integrated Data Office, Ministry of Jobs, Trade & Technology, Government of BC

[3]Population Data BC, University of British Columbia

Introduction

Ligo is an open source application that provides a framework for managing and executing administrative data linking projects. Ligo provides an easy-to-use web interface that lets analysts select among data linking methods including deterministic, probabilistic and machine learning approaches and use these in a documented, repeatable, tested, step-by-step process.

Objectives and Approach

The linking application has two primary functions: identifying common entities in datasets [de-duplication] and identifying common entities between datasets [linking]. The application is being built from the ground up in a partnership between the Province of British Columbia's Data Innovation (DI) Program and Population Data BC, and with input from data scientists. The simple web interface allows analysts to streamline the processing of multiple datasets in a straight-forward and reproducible manner.

Results

Built in Python and implemented as a desktop-capable and cloud-deployable containerized application, Ligo includes many of the latest data-linking comparison algorithms with a plugin architecture that supports the simple addition of new formulae. Currently, deterministic approaches to linking have been implemented and probabilistic methods are in alpha testing. A fully functional alpha, including deterministic and probabilistic methods is expected to be ready in September, with a machine learning extension expected soon after.

Conclusion/Implications

Ligo has been designed with enterprise users in mind. The application is intended to make the processes of data de-duplication and linking simple, fast and reproducible. By making the application open source, we encourage feedback and collaboration from across the population research and data science community.

http://dx.doi.org/10.23889/ijpds.v3i4.749