MLDS Center
Maryland Longitudinal Data System

Address 550 West Baltimore Street
Baltimore, MD 21201
Phone 410-706-2085
Email mlds.center@maryland.gov
Website www.MLDSCenter.org

Monthly Report for August 2015

Submitted September 2, 2015

Staffing

An offer was made to an individual for the network administrator position. However the offer was turned down and staff will readvertise the position.

Meetings

Legislative Briefing

The *Joint Committee on Cybersecurity, Information Technology, and Biotechnology* held a briefing on the use of big data to assist government in providing public services in law, education, human resources and transportation. Ross Goldstein and Dawn O'Croinin attended the hearing. In general, the presenters talked about the various ways to use data to better inform policies and improve processes. There was also a broad acknowledgement of the challenges of collecting and integrating data from disparate systems. The education presenters were from University of Maryland, University College and discussed how they used data for enrollment management, trend identification and generally to improve classroom practice.

Dual Enrollment

Ross Goldstein attended a meeting on dual enrollment reporting with Jack Smith, Henry Johnson, and Chandra Haislet from MSDE, Brad Phillips from the Maryland Association of Community Colleges, and Jon Enriquez from MHEC. The discussion focused on the upcoming reporting requirements and the quality of the data available. The Center is required to submit an annual report to the General Assembly on December 15th on the number of students dually enrolled and the classes in which they are enrolled. MSDE initiated a new data collection to identify students dually enrolled and their courses for the 2013-2014 academic year. MSDE has analyzed that data and has determined that there were reporting problems and omissions by several LEAs. MSDE will work with the LEAs to improve the reporting for the 2014-2015 academic year. The Center agreed to provide its initial findings at a subsequent meeting in late September or early October. This will allow all parties to understand the data and provide a quality review.

System Development

Completed Tasks

Staff has completed the installation of Embarkedero (data architect software). Staff is in the process of giving users access to data models and data dictionary.

Staff has also completed the installation of MFT (master file transfer) and relocated files and folders from DPSCS. Staff upgraded MFT version to include PGP encryption for the files and is in the process of migrating users and setting up configuration for secure file transfer.

Staff completed an *Incident Management Plan* to guide the planning and response to an IT security incident. A copy of the plan is attached. Finally, staff has completed all requirements for Universal License Agreement certification with Oracle.

Upcoming Tasks

Staff will be working towards:

- Finalizing the firewall setup with DoIT;
- Setting up a permanent offsite backup solution; and
- Upgrading all components of Oracle (Oracle Database, Webcenter and OBIEE), hardening all passwords, completing implementation of VNX upgrade to 5700, and migrating from OWB to ODI.

Data Loading

Staff has made a lot of progress in loading data from the agencies.

1. MSDE - all attendance, assessment, and National Student Clearinghouse data has been loaded. Staff is in the process of loading the SCGT (Student, Course, Grade, Teach) data. A new National Student Clearinghouse File is also being loaded.
2. MHEC - data from the Degree Information System (DIS), Enrollments Information System (EIS), High School Graduate System (HGS), Financial Aid Information System (FAIS) (2008-2012), has been loaded. Center recently received the MAC2 collection for EIS and DIS and are in the process of loading those files.
3. DLLR - all data from DLLR (Wage data, UI Tax, NEDP, LACES, and GED) has been loaded and the Center is now loading the last quarter's data.

The following table shows the summary counts of data loaded into the system from the partner agencies.

Data Source	Distinct count	Identifier Counts		
		SASID	Valid SSN	Name and DOB
K12 (includes teachers)	1,481,447	1,321,387	1,004,374	1,436,706
MHEC	873,744	182,730	873,739	467,652
DLLR	4,720,963	370,539	4,707,692	1,273,851
Net Total	6,044,232	1,348,106	5,551,162	2,461,644

As of 8/20/2015

Data Challenges

Background

For all data related to a person, MLDS assigns a synthetic ID whose value has no resemblance to the actual personal identifiers contained in source data. Such an identifier is often referred to as a "research ID" or "RID".

Each RID assigned in the MLDS system should represent a separate individual. However, due to the difficulty of reconciling conflicting personal identifiers received from multiple data sources, MLDSC uses "deterministic matching" when new data is loaded. This requires an exact match of personal identifying ("PII") data in order for newly-imported data to be assigned to an existing ID. For example, the MLDS system would assign 3 separate RID's for these sets of fictitious PII values:

Data Source	SASID	SSN	DOB	Last	First	Middle
MHEC-EIS	-	123456789	2/3/1998	LOPEZ	JEREMIAH	OCHOA
MSDE-SCGT	9876543210	-	3/2/1998	OCHOALOPEZ	JEREMIAH	-
DLLR Claims	-	123456798	2/3/1998	OCHOA	JEREMIAH	-

Issue

As a consequence there are many instances where one or more RID's have been assigned to the same person. The Center has a process for using "probabilistic matching" to identify highly likely duplicates, and to date just over 200,000 sets of separate identities have been merged into single identities using this process.

However, staff has reached the practical limit of what can be done to identify duplicate identities using only MLDSC internal data.

Solution

Verifying personally identifiable information using data from a third party source would provide a solution for resolving identities and enabling further linking of data. Dawn O'Croinin is working with MVA to determine whether its data can serve as a source for verification.

The following are examples of where third party identity verification could be effectively utilized.

1. *High school graduates with missing SSN's* - MLDS contains records for 470,000 students who exited 12th grade in school years 2008-2014. About 75,000 of these (17%) are without a valid SSN in the system. The identities for these students cannot be related to wage data, because SSN is the only PII identifier supplied for wage data received from DLLR.
2. Multiple SSN's for the same SASID (State-assigned student ID) - MLDS contains about 3,000 identities (1% of 12th grade students) associated with more than one SSN. Third-party verification can be used to verify which of the conflicting SSN's actually belongs to the student.
3. A single SSN is associated with multiple names or dates of birth - MLDS contains about 16,000 identities (3% of 12th grade students) associated with a single SSN, but having more than one first/last name combination or date of birth. Third-party verification can be used to verify which of the names and dates of birth are correct for the SSN, and whether the alternate name and date of birth can be associated with another SSN.
4. DLLR wage record matches - DLLR wage records include no personal identifiers other than an SSN and a 4-letter string consisting of first initial and first 3 letters of the last name. About 1,252,000 of the SSN's from DLLR wage SSN's match with the SSN's from MLDSC's other data sources (MSDE and MHEC). For the large majority (98%) of the SSN matches, the DLLR initials match the first/last names provided by MSDE, but there are about 26,000 SSN's where the DLLR initials do not match with the MSDE names. Third-party validation will be helpful in resolving whether these SSN's actually belong to the identified students.

Data Management

Laia Tiderman has completed the Data Collection Calendar for the 2015-2016 academic year. The calendar was presented to the Data Governance Advisory Board for input. The Data Collection Calendar will be presented at the September Board Meeting for approval.

Research

The 2012 SLDS Grant required MSDE to conduct "training academies" for legislators and legislative staff. Pursuant to an agreement, the Center will perform the trainings. The first training will be for 5-10 legislative staffers later this month. The Research Team will conduct the training and cover various topics including an in-depth understanding of the system and the data collected, important concepts to understanding large longitudinal data sets, limitations of the data, the research methodologies to be used, and initial findings from research conducted to date. The session will be for 5 to 10 staffers and be for two to three hours.

The Research Team is also making continued progress on various reports including:

- Assessing STEM post-graduate student state and regional job acceptance and retention;
- Impacts on students assigned to remedial courses in postsecondary education; and
- Early Childhood Workforce Retention.