Accurately matching data for individuals as they move through different stages of education and employment is the key to creating longitudinal datasets and to building a successful statewide longitudinal data system (SLDS). The match rate—or the level of accuracy with which a state can identify records for the same individual across multiple agencies—is an important indicator of the quality, completeness, and usefulness of longitudinal data. The match rate depends on a variety of factors. Data quality; the matching technology in use; human, technical, and fiscal resources available; and the needs and priorities of partner agencies all play a role in determining and achieving an acceptable match rate for the state’s data use needs.

SLDS leaders from Texas and Hawai’i discuss their methods and tools for matching records across agencies, the match rates they have achieved, and how they use the resulting longitudinal data. Additionally, Connecticut’s SLDS program manager shares that state’s approach to evaluating and improving its match rate.

Texas: Match Rate as a Key Part of Transparent Data Use

Texas has been linking public K12 and postsecondary data since 2000, when the Texas Education Agency (TEA) and the Texas Higher Education Coordinating Board (THECB) signed a data sharing agreement allowing them to share personally identifiable information for students. The agreement paved the way for the creation of a P-16 (early childhood through postsecondary) SLDS that also includes school employment data for educators.

Data matching process

TEA and THECB each can match the other’s records with its own as needed using students’ names, dates of birth, and Social Security numbers, which are collected by TEA at the K12 level as well as by THECB. Using these data with occasional variations—such as matching the first three letters of students’ first names rather than full names—the agencies are able to find about 90 percent of college-bound public K12 graduates in public and private colleges and universities in Texas. The state’s large student population makes it cost prohibitive to obtain National Student Clearinghouse data every year for students attending out-of-state postsecondary institutions. However, special projects involving National Student Clearinghouse data in the past have confirmed that about 10 percent of Texas public high school graduates enroll in out-of-state postsecondary institutions.

For the Texas public K12 students who do not have or provide Social Security numbers, the state is encouraging higher education institutions to store the state K12 identifier available on high school transcripts and report it to THECB to facilitate data matching. Students for whom K12 records cannot be matched to postsecondary records are classified as either “not found”—meaning the state has a valid identifier for the student but could find no record of postsecondary enrollment—or “not trackable”—meaning the state does not have an identifier for the student. Rates of untrackable students range from 2 percent to about 15 percent in each K12 district, with a statewide rate of about 5 percent.

This product of the Institute of Education Sciences (IES) was developed with the help of knowledgeable staff from state education agencies and partner organizations. The content of this publication was derived from a Statewide Longitudinal Data Systems (SLDS) Grant Program monthly topical webinar that took place on September 16, 2015. The views expressed do not necessarily represent those of the IES SLDS Grant Program. We thank the following people for their valuable contributions:

**Webinar Presenters:**

- Julie Eklund  
  Texas Higher Education Coordinating Board  
  Anita Huang  
  Hawai’i P-20 Partnerships for Education  
  Todd Ikenaga  
  Hawai’i P-20 Partnerships for Education  
  Jan Kiehne  
  P20 WIN, Connecticut Board of Regents for Higher Education

**Moderator:**

- Hans L’Orange  
  SLDS Grant Program, State Support Team

For more information on the IES SLDS Grant Program or for support with system development, please visit [http://nces.ed.gov/programs/SLDS](http://nces.ed.gov/programs/SLDS).
Using matched data

Texas uses its matched K12 and postsecondary data to satisfy state and federal reporting requirements, including those for state performance reports, EDFacts, Career and Technical Education performance measures, and Elementary and Secondary Education Act (ESEA) Flexibility. It also publishes postsecondary outcomes reports, many by school district and campus, including dashboards featuring higher education enrollment, student grade point averages, developmental education enrollment, and related college readiness information of interest to K12 districts and the public. An agreement with the Texas Workforce Commission allows THECB to link data for high school and postsecondary graduates with wage data for further insight into how students fare in the workforce.

Additionally, THECB provides linked K12 and postsecondary data to up to three Education Research Centers (ERCs) housed and sponsored by public colleges or universities, which can provide the data to researchers for approved research purposes. Student-level data provided to the centers are stripped of personally identifiable information and, as such, are accessible for use at the ERC facilities for matching across sectors. Researchers’ results are only approved for release if the data are aggregated and cell values representing small numbers of individuals are suppressed to comply with Family Educational Rights and Privacy Act (FERPA) requirements.

Match rate improvements and lessons learned

TEA and THECB make an effort to communicate clearly and transparently with their stakeholders about how K12 and postsecondary data are used and matched. Match rates are published alongside many data reports to clarify how many individuals the report covers and how many might be missing. To help improve the match rate, THECB developed a process for colleges and universities to correct or update Social Security numbers and other student identifiers that were omitted or incorrectly reported to the board in the past.

The agencies’ efforts to justify each piece of data they collect and to ensure that data are used appropriately and securely have helped them build high levels of trust with one other and with stakeholders over time. This trust, along with matched data made available to a range of education and workforce stakeholders, have led to continued support for and reliance on SLDS data at the state level. The state legislature has called for continued data sharing across agencies, and the state’s 15-year strategic plan has set new data-driven targets for higher education enrollment.

Hawai’i: Adapting Processes to Fit the Data

Hawai’i P-20 Partnerships for Education (Hawai’i P-20) manages the state’s P-20W (preschool through workforce) SLDS from within the University of Hawai’i. Hawai’i P-20 is responsible for collecting and matching K12 data from the Hawai’i State Department of Education (DOE), postsecondary data from the University of Hawai’i System, and unemployment insurance data from the state Department of Labor and Industrial Relations. The state’s small size and centralized education structure help streamline data matching between the K12 and postsecondary sectors. DOE operates as a single statewide school district, making it the central source for public school data, and the University of Hawai’i System gathers data from all of the state’s public colleges, universities, and community colleges into a single student information system. About two-thirds of Hawai’i’s college-bound high school graduates enroll in one of the university system’s 10 campuses.

Data matching process

Beginning in 2012, Hawai’i P-20 developed a series of deterministic rules for matching individual records across sectors and agencies. These rules use a combination of data—including first name, last name, date of birth, and gender—and exact and partial matching criteria to identify records pertaining to the same individual. Each rule is evaluated for consistency, and rules with very high rates of accuracy are used to match cross-sector data automatically. If a rule yields false positives (matched records that actually belong to different individuals) or false negatives (failures to match records belonging to the same individual), all records

Hawai’i P-20’s Deterministic Matching Rules

Records are matched automatically under the following rules:

1. Exact last/first name, date of birth, and gender
2. Exact last/first name and date of birth
3. Exact last name, date of birth, and gender; embedded first name
4. Exact first name, date of birth, and gender; embedded last name
5. Exact date of birth and gender; embedded first and last names
6. Exact last name, date of birth, and gender; first three characters of first name
7. Exact last name and gender; first three characters of first name; birth year and month

Additional matching rules are used to identify possible matches that are reviewed manually by Hawai’i P-20 staff.
matched using that rule are reviewed manually by Hawai'i P-20 staff.

In 2013, Hawai'i P-20 purchased third-party vendor products to implement its SLDS, including a record-matching solution that used both deterministic and probabilistic methods. The vendor's deterministic matching engine used first name, last name, and date of birth to link records, but the number of false positives and false negatives was higher than Hawai'i P-20 liked.

The probabilistic engine also used first name, last name, and date of birth to find matches, and it automatically merged records meeting a certain threshold of similarity. Once these records were merged, Hawai'i P-20 staff members could not easily override matches between records that actually represented two different people. Adding to the system's challenges, Hawai'i P-20's attempt to process 10 years of historical records at once resulted in 25,000 possible matches requiring manual review—an overwhelming number for the SLDS staff.

Apart from adding a process for identifying twins—which often cause false positive matches due to their similar data—the vendor was unable to offer improvements to its matching processes to decrease the error rate. Hawai'i P-20 decided to augment the vendor's process by running records through its own deterministic rules first. The agency was able to match and split records according to its own criteria before loading them into the vendor's system, which reduced the number of historical records requiring manual review to 10,000. Processing the historical data chronologically by school year or postsecondary term rather than in one large batch also improved the state's match rate. Hawai'i P-20 staff members still review all records matched using the vendor's probabilistic method, but this process is manageable due to the state's relatively small population.

Using matched data
One of Hawai'i's first goals for its matched, longitudinal K12 and postsecondary data was to revamp an existing College and Career Readiness Indicators report, which provides information about students' enrollment in and preparation for higher education and employment. When the report was introduced in 2008, it used data gathered and aggregated separately from each sector and relied on student-reported postsecondary enrollment information. Beginning with the high school graduating class of 2012, Hawai'i has used its matched records to provide more accurate information about college and career readiness. Hawai'i P-20 has committed significant time and resources to maximizing its match rate in order to provide the most accurate information possible for this and other products.

Match rate improvements and lessons learned
The limitations of the vendor-designed match process reinforced for Hawai'i the importance of knowing its own data and finding ways to adapt the matching process to fit those data. By supplementing the vendor's tool with matching rules it developed and tested in-house, Hawai'i P-20 has improved its match rate. Continued improvements such as a twin indicator, manual merging and splitting capabilities, and batch processing of historical data help refine the process to fit the state's data.

Connecticut: A Closer Look at Match Rate

From the start of Connecticut's SLDS grant work in 2009, project leaders for the state's Preschool through 20 and Workforce Information Network (P20 WIN) planned to evaluate the effectiveness of their record-matching process. P20 WIN links public K12 data from the Connecticut State Department of Education; postsecondary data from the Connecticut Board of Regents for Higher Education—which oversees the public state universities, community colleges, and an online institution—the University of Connecticut, and the Connecticut Conference of Independent Colleges; and workforce data from the Connecticut Department of Labor. Project leaders realized that demonstrating an accurate and reliable match rate would be key to sustaining the support and confidence of these partner agencies, which participate in the network voluntarily rather than through a state mandate.

Evaluating prospective matching tools
Connecticut tested and evaluated data-matching tools in order to select a data-matching utility for its SLDS. State leaders knew the chosen method would need to reliably match K12 and postsecondary data probabilistically based on name, date of birth, high school, and a non-universal student identifier, as well as match student data to workforce records based on Social Security number and partial names. P20 WIN’s project manager and data stewards developed a test plan to evaluate both in-house and vendor-produced matching tools using fabricated and modified real-life data.

Each matching tool was run through a series of test cases to reflect a variety of realistic data-matching scenarios. The test cases were designed to assess how each tool handled datasets of different sizes and data with modified or missing values. Although there were no “right answers” when matching fabricated data, the P20 WIN team calculated Quality Match Ratios based on a logical evaluation of each tool's output. The ratios and additional evaluation criteria, including the tool's usability and data-processing capacity, informed Connecticut's decision to purchase a matching tool from a vendor.
Calculating and evaluating the match rate

With the selected tool in place, P20 WIN’s partners decided to evaluate its match rate compared to a baseline of matched data from the National Student Clearinghouse. P20 WIN sent K12 data for a cohort of more than 38,000 students to the National Student Clearinghouse, which returned matches for about 15,500 students enrolled in public colleges and universities in Connecticut. When the same K12 data were linked to postsecondary data from the Board of Regents using P20 WIN’s matching tool, nearly 16,500 matching records were returned. P20 WIN consulted with the National Student Clearinghouse about its matching process and determined that the differing match rates were likely due to the Board of Regents having more recent data and no method for students to opt out of the data collection, as well as differences in matching methodology.

The Connecticut Department of Labor conducted a further examination of records considered to be low-probability matches, such as those with probabilistic scores around 30 percent. Investigators found no obvious mismatches using the matching tool even at these low thresholds, prompting the Board of Regents to conduct a manual review of selected records to get a closer look at the match rate. In a sample of 1,000 matched records, reviewers found two poor matches and three more that were questionable, indicating a match rate of over 99 percent. Even if actual match rates for records over time are somewhat lower, the evaluations left P20 WIN’s leaders and partners with a high level of confidence in the matching tool’s effectiveness.

Next steps

P20 WIN plans to continue evaluating its matching tool as new and different data are incorporated into the SLDS. The partners are currently exploring methods of matching K12 data to Department of Labor data using state Division of Motor Vehicles records to supplement the limited name data available in workforce records. When these records are in use, P20 WIN will assess the match rate with a particular focus on identifying any biases in the dataset that might impact education program evaluations performed with SLDS data.

Additional Resources

Connecticut P20 WIN
http://www.ct.edu/initiatives/p20win

Hawai'i P-20 Partnerships for Education
http://www.p20hawaii.org/

Linking Early Childhood and K12 Data: A State Example from Kentucky: SLDS Webinar
https://slds.grads360.org/#communities/pdc/documents/6948

Linking K12 Education Data to Workforce: SLDS Webinar
https://slds.grads360.org/#communities/pdc/documents/5871

Linking K12 Student Data with Postsecondary Data: SLDS Webinar
https://slds.grads360.org/#communities/pdc/documents/5793

Texas Education Agency
http://tea.texas.gov/

Texas Higher Education Coordinating Board
http://www.thecb.state.tx.us/

Using DMV Records to Access Social Security Numbers: SLDS Webinar
https://slds.grads360.org/#communities/pdc/documents/5909