IES Institute of Education Sciences

Interagency Data Linking and Common Identifiers

SLDS ISSUE BRIEF
June 2020

U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Statistics at IES

This product of the Institute of Education Sciences (IES) Statewide Longitudinal Data Systems (SLDS) Grant Program was developed with the help of knowledgeable staff from state education agencies and partner organizations. The content of this publication was derived from an SLDS Personnel Exchange Network meeting held March 9, 2020, in Sacramento, California. The information presented does not necessarily represent the opinions of the IES SLDS Grant Program.

For more information on the IES SLDS Grant Program or for support with system development, please visit *http://nces.ed.gov/programs/SLDS*.

CONTRIBUTORS
Kathy Gosa and Jeff Sellers, *SLDS Grant Program State Support Team*
Meredith Fergus, *Minnesota Office of Higher Education*
Chuck Murphy, *Kentucky Center for Statistics*
Tim Norris, *Washington State Education Research & Data Center*
Courtney Petrosky, *Minnesota IT Services*

Developing high-quality longitudinal data to trace long-term education and employment outcomes and inform state policy and practice depends on accurately matching information about specific individuals across multiple data sources. Over time, P-20W+ (early childhood through workforce) statewide longitudinal data system (SLDS) programs need to adapt and improve their data linking strategies as they incorporate additional years of data and data sources into their systems and modernize their information technology (IT) infrastructure.

This brief offers best practices and strategies from several states related to setting goals for cross-agency data sharing, identifying requirements from data sharing partners, establishing technical solutions and processes to link data in a P-20W+ SLDS, and collaboratively governing interagency data collections.

Data Linking Goals and Partners

The data incorporated into a P-20W+ SLDS reflect the information that state agencies need to ensure accountability, evaluate programs, inform decisions and practices, and improve policies. Similarly, the processes and systems used to link data from different sources within the SLDS must reflect the goals and needs of data contributors and the state as a whole. When designing data linking strategies, SLDS teams need to gather input from relevant partner agencies and stakeholders, define the SLDS's goals and data needs based on that input, identify which of its partners' data will be needed, and determine where and how to obtain those data.

The following guiding questions can help SLDS teams develop their approaches to interagency data linking:

- Which SLDS team members and stakeholders should be involved in various phases of the data linking process?
- How will security and legal requirements for specific types of data be addressed?
- How will education data be linked to data from other sectors, such as health, human services, and workforce?
- When data are not readily available from an SLDS partner, such as private university data, how will the SLDS team identify the best sources for those data?
- How will ongoing quality of the data linking be measured, reported, and managed?

State highlight: Minnesota's enhanced data linking project

Minnesota state agencies collaborate on two interagency data systems. The Statewide Longitudinal Education Data System (SLEDS) integrates data from K12 and postsecondary education agencies as well as workforce and employment data. The Early Childhood Longitudinal Data System (ECLDS) contains data from a variety of early childhood care and education programs serving children from birth through third grade. Both data systems share IT infrastructure managed by Minnesota IT Services (MNIT).

With the state's 2015 SLDS grant, MNIT began exploring ways to improve and enhance its data linking processes. The initiative emerged out of growing interest among the state's early childhood and education partners in how factors such as parental education, earnings, and marital status affect young children's outcomes. Answering these questions would require not only linking data for individual children across multiple state programs and services, but also establishing relationships between those children and their parents' data.

MNIT conducted a data profiling exercise with its SLEDS and ECLDS data contributors to review and document current data sharing processes. IT representatives gathered a standard set of information about each data source, including names and descriptions of the data files being shared and expected formats and values for specific data elements. Following the exercise, MNIT established standard protocols for data submissions, including file naming conventions, to ensure a consistent submission process. The exercise also allowed MNIT to determine the data elements most commonly used to identify individuals in each data source and how to most effectively link data records across those sources.

To improve its data linking technology, MNIT interviewed SLDS teams in 13 states to learn which tools they used to link data. The agency then contacted vendors for those tools for more information and demonstrations, and it engaged all state agencies involved in SLEDS and ECLDS to review the vendor-provided information. With the agreement of its partners, MNIT issued a request for proposals to select a new set of tools for data linking, data profiling, and reporting and reviewing data matching errors.

As the new tools are implemented, a workgroup of stakeholders is reviewing the initial results of relationship-based data linking and creating policy questions that the state will try to answer with its enhanced data links.

Technical Solutions and Processes

The technology used to link P-20W+ SLDS data and the processes supporting it need to reflect the goals, objectives, and technical capacity of the data sharing partners and data users. The following guiding questions can help SLDS teams design or choose data linking tools that fit their data system environments and needs:

- Does the SLDS use a centralized, federated, or hybrid data system model? For more information about the structure and governance of different data system models, see *SLDS Issue Brief: Centralized vs. Federated: System Models for P-20W+Data Systems* (*https://slds.ed.gov/#communities/pdc/documents/17542*).
- Will the data be hosted on cloud-based servers or on-premise servers?
- What are the sources and formats of the data being shared?
- What is the schedule and frequency for importing and matching data in the SLDS?
- What identifying data elements are available from data contributors?
- What matching rules, business rules, and processes will be used to link data?
- What resources—including staff capacity—are available for linking data?
- Will the data linking solution be built by the SLDS agency, or will a vendor product be used? What are the tradeoffs of in-house versus vendor solutions?

Minnesota's SLDS infrastructure and data matching

Minnesota's SLEDS and ECLDS are centralized systems in which data from seven contributing state agencies and programs are integrated into dedicated warehouses. Data from partners' source systems are loaded twice a year via a secure file transfer protocol (FTP). Identifiable data, including names, Social

Security numbers, and student IDs, are stored on an MNIT file server where only authorized MNIT staff members can access them. De-identified individual records are moved to an operational data store and assigned a Reported Person ID. For additional security, new Reported Person IDs are generated with each 6-month data import so that individuals' records cannot be re-identified from one data load to another.

MNIT uses individuals' names, dates of birth, Social Security numbers, and K12 student IDs to link data records from different source systems. The agency's linking algorithms use seven matching rules to compare identifying data elements across records and determine which records belong to the same individual. Once linked, data from the operational data store are moved to separate warehouses for ECLDS and SLEDS, where they are used for public reporting and analysis.

MNIT uses SAS applications for data linking and workflow management, replacing an older matching system that was built within the department. The team continues to enhance its new applications to improve efficiency and automate initial error checks when datasets are uploaded.

Minnesota's SLEDS and ECLDS are staffed by 10 full-time IT employees and 2.5 full-time equivalent positions for management and communications. Each data contributing partner agency designates a coordinator who participates in data governance and is responsible for data transfer and validation. The two data systems receive $2 million each year in state funding through the Minnesota Office of Higher Education.

Washington's SLDS infrastructure and data matching

The Washington State Education Research & Data Center (ERDC) maintains a centralized P-20W+ SLDS within the governor's Office of Financial Management. The SLDS contains data from early childhood, K12 and postsecondary education agencies, and workforce programs along with some additional state services.

The schedule and format for submitting data to the ERDC depends on the source. Sources submit their data files annually, semiannually, or quarterly in a format agreed upon with ERDC in advance. Data submissions are pulled from a source staging area, loaded into a pre-stage database platform, and then moved into a staging database where they undergo quality and validation checks. From there, necessary identifiers are moved into ERDC's master data

Considerations for assigning unique identifiers

Unique identifiers (UIDs) associate data about the same individual from different records or sources within a P-20W+ SLDS. UIDs allow for easier matching, storing, and sharing of data across systems and can take the place of personally identifiable information like names and Social Security numbers. When developing a system of assigning and using UIDs in an SLDS, consider the following questions:

- *What identifiers are already used in state data systems?* Gather information from each data contributor about the identifiers they currently use. Determine the quality of these data elements and consider how they might be used to link individual records under a common P-20W+ identifier.
- *What is the overall design of the SLDS?* Characteristics such as whether the SLDS uses a centralized or federated system model can help determine aspects of a UID system that will be effective or practical.
- *What privacy laws might impact the SLDS?* Examine existing data privacy laws that govern the agencies and programs that contribute data to the SLDS. These laws include federal regulations such as the Family Educational Rights and Privacy Act (FERPA) and the Health Insurance Portability and Accountability Act (HIPAA), as well as state data privacy laws. Consult with partner agencies' legal representatives to clarify privacy considerations for the SLDS.
- *How will personally identifiable information be protected?* UIDs need to be protected through secure authentication processes for SLDS users, with access to identifying information granted only to appropriate individuals.
- *How will near matches be reviewed and handled?* To effectively manage UIDs, SLDS programs must establish data governance processes for reconciling individual records that cannot be linked with certainty by matching rules and algorithms. Tools and reports may be used to support the data governance and review processes for designated staff members to manually review near-match records to determine whether to join them under a single UID or assign a new UID.

management system for cross-sector linking and then into an operational data store and datamarts for research and fulfilling data requests.

ERDC asks its data contributors to share any identifying data elements they collect that could help match individual records with other sources. Incoming data records are assigned a temporary unique identifier known as a "Pkey" to identify them in the context of the source dataset. In K12 datasets, for example, Pkeys identify unique students within a school district. A student who transfers and has records in two districts will have two Pkeys. ERDC's identity resolution engine applies several levels of matching algorithms to link records across datasets based on identifying data elements such as name and Social Security number. ERDC also may use other information, such as monthly enrollment patterns or first date of employment, to help corroborate individuals' identities.

ERDC's matching process errs on the side of undermatching, or treating data records with similar identifiers as separate individuals until they can be confirmed as a single person. The center maintains a "green list" and "red list" of individuals whose data either should or should not be merged regardless of the automated matching results. ERDC reviews its matching rules annually for potential improvements.

ERDC uses Informatica software to manage its data matching process. The center is staffed by 10 researchers, half of whose positions are grant funded, along with 2 grant-funded data warehouse developers, 3 data warehouse maintenance staff, and a full-time identity resolution manager. Each data source has a dedicated data steward and custodian at ERDC, but source agencies do not provide additional staff or support to ERDC. ERDC's data warehouse hosting services are moving from Washington's state IT agency to the Office of Financial Management's internal IT division.

Kentucky's SLDS infrastructure and data matching

Kentucky's centralized P-20W+ longitudinal data system is managed by the Kentucky Center for Statistics (KYSTATS), an independent state agency that collects and integrates data from 12 partner agencies and programs.

Data contributors can submit files via the KYSTATS website, a secure FTP, a web service, or by allowing KYSTATS to pull data directly from their databases. Each contributor has a dedicated submission window during the year, and KYSTATS schedules system updates around these windows to avoid interfering with data transfers. Once submitted, data files undergo an initial validation that catches null records and unexpected values. A second validation happens after incoming data are linked to existing records. Preprocessing, matching, and exception handling occur in a data staging area, then personally identifiable information is stripped, and records are stored in a de-identified data reporting system.

KYSTATS uses its data contributors' preferred identifiers for linking internally, and source identifiers

State highlight: Kentucky's identity resolution process and quality control

The Kentucky Longitudinal Data System (KLDS) uses a Master Person Index to save a "golden record" of the most recent data for each individual in the system. Each Master Person Index record in the KLDS can have multiple "alias" records to store alternate data, such as previous names.

To match data records, KYSTATS uses an agency-developed identity resolution engine that applies matching rules that are customized to the data source's preferred identifiers and data quality limitations. Progressively "fuzzy" levels of matching rules are applied to account for misspelled names or other data errors that might prevent data belonging to the same individual from matching under stricter algorithms. KYSTATS analysts can manually review data that are unmatched or questionably matched to other records through an interface in the center's data matching system. Analysts can combine or split data from records if they determine the data do or do not belong to the same individual. All record changes are logged to help identify common matching issues and trends.

KYSTATS reviews samples of data records matched at each level of identity resolution to assess the accuracy of each matching rule. Over time, matching rules may be retired if different data elements become available or changes in data quality make them less useful.

are mapped to standardized element names to ensure comparability between systems. KYSTATS's matching algorithms are different for each data source. Each newly imported record is assigned a point-in-time identifier (PID). After matching, data are linked to existing SLDS records with a global unique identifier (GUID) or assigned a new GUID.

KYSTATS's data warehouse and matching systems were developed internally, with some public reports built in Tableau. KYSTATS staff include an executive director; 8 developers; 10 analysts, data scientists, and visualization experts; 6 business and operations staff members; and 13 staff members from Kentucky's former Labor Market Information Office that merged with KYSTATS. KYSTATS pays to host its systems on servers owned by the central state IT agency.

Interagency Governance for Data Linking

Data governance is the means by which organizations make decisions about their collective information assets. An interagency data governance program oversees the integration of data across organizations based on agreed-upon business rules. It is also the means by which data contributors collectively ensure appropriate use of linked data.

The following guiding questions can help interagency data governance teams plan for data linking:

- Who will make decisions about developing the matching process, and how will changes be made to the process?
- Will external research designs and reports that use linked data be shared with the data contributors?

Interagency data governance programs can support the effective and responsible use of linked data in three broad areas: creating data access and management policies, developing research or data request processes, and establishing processes for incorporating new data and data-contributing partners.

Data access and management policies

Data access and management policies outline who will be able to view and use interagency SLDS data and how the data will be maintained and used. Creating these policies requires strong, consistent communications and trust among partners that the data they contribute to the SLDS will be used responsibly. In Minnesota, the SLEDS partners hired a third-party organization to draft its data management and use policy. The neutral facilitator and 2.5-year development process helped ensure a thorough policy that had the confidence of all partners. The policy aligns with state data management regulations and sets protocols for approving data uses and ensuring the confidentiality of published data. It also outlines six levels of access for data users, ranging from highly restricted access for approved MNIT staff to public datasets and reports.

Research and data request processes

Formal processes for requesting SLDS data help put linked, interagency data in the hands of researchers who can help inform policy and evaluate state programs. Research request processes—also called data request processes—ask potential data users to specify the data they plan to use and how, and they establish criteria and processes for interagency data governance groups to review and approve requests. Many SLDS programs use a standardized application form and review rubrics to collect and evaluate data requests. In interagency data governance programs, each partner agency whose data are involved in the request may need to sign off on the request. Some SLDS programs establish separate research request processes for requests coming from a data contributing agency rather than from an external

State highlight: Washington's policy inventory and matrix

To keep up with evolving security issues, ERDC's data management system includes extensive documentation of policies and processes surrounding all datasets coming to the SLDS from contributing partners. A policy inventory and matrix record the requirements, data sharing agreements, and processes affecting each dataset and technology asset. The matrix identifies the organization responsible for the data, whether it is the state IT agency, the Office of Financial Management's IT department, or ERDC's IT team. It also documents a set of internal controls and monitoring procedures including the frequency, location, and signoff required for security checks.

Based on the policy inventory, ERDC identified and addressed gaps in existing policies and gained leadership approval for all policies surrounding SLDS data. The policy inventory is stored on an internal SharePoint website, and an individual is assigned to update it as new policies come into place.

researcher. Different processes also may be used for requests involving individual-level data versus aggregated, de-identified data.

Incorporating new data or data contributors

Interagency data governance programs need to agree on how new datasets or data from new partner agencies will be added to the SLDS. In Minnesota, the SLEDS data governance team asks potential data contributors a set of standard questions to assess their capacity and readiness to join SLEDS (see box at right).

As new data sources come on board, the technical processes for data integration need to be formalized. In Washington, ERDC research staff members meet with representatives from the data contributor to confirm the data elements, option sets, and business rules associated with their data. A formal process of data readiness, profiling, and loading standardizes the source data for easier integration into the SLDS. KYSTATS holds similar discussions with new data contributors to establish which data will be shared, when data will be submitted, and how to ensure data quality.

Conclusion and Lessons Learned

P-20W+ SLDS programs must continually reevaluate and modernize their data system infrastructure and linking processes to accommodate evolving data collections, meet emerging needs, and take advantage of new technology. By evaluating past work, planning for next steps, and learning from similar programs in other states, SLDS teams can improve their systems and linking strategies.

SLDS teams in Kentucky, Minnesota, and Washington offer the following lessons learned from their experiences with interagency data linking.

Data linking is part art and part science

P-20W+ SLDS programs do their best to develop strong data linking algorithms that can compare and accurately match information about the same individual from multiple data sources. However, no algorithm is fool proof. By regularly reviewing the results of data linking, SLDS teams can improve their systems to recognize and address common errors, eliminate ineffective matching rules, and account for differences in record keeping data quality across agencies.

Minnesota SLEDS questions for new data contributors

- Is your agency willing to engage in this data sharing partnership? Are your leadership and stakeholders on board?
- Can you dedicate staff time and resources to the partnership?
- Does your agency have legal authority to share data? Are there statutory conditions to consider?
- How are the data to be shared collected and stored? Can you pull and test the data regularly for reliability?
- Have you considered how to use SLEDS data within your agency?

Plan for staffing needs and training

Whether data linking management processes are developed by an internal IT team or a vendor-supplied tool, the SLDS team needs personnel who understand the interagency datasets and who can effectively administer and support the data governance processes related to linking. These staff members will need sufficient time to build their capacity to understand the data as well as the tools, particularly when implementing new systems.

Plan for system development and upkeep

As P-20W+ SLDSs are built and implemented, agencies often find that they need features, architecture, or data sources outside of their original plans. Additionally, maintaining and enhancing an SLDS to remain useful and relevant for stakeholders can cost as much as its initial development. In many cases, state budget funds will need to replace grants to sustain the system.

Secure support from leaders and stakeholders

Successful interagency data systems depend on state and agency leaders and stakeholders buying in to the system's value and committing to share data. Helping stakeholders understand how their data will be integrated and managed establishes transparency and strong partnerships. In Minnesota, ECLDS leaders include a chart illustrating the system's data integration workflow in data sharing agreements to help gain the participation of new partners.

Make time for review

Testing data linking results and monitoring match quality through error dashboards and regular reviews helps improve the linking process and saves time for future datasets imported into the SLDS. It is important not only to address current linking issues, but also to build in flexibility to anticipate future challenges.

Additional Resources

California Data System Common Identifier Background Paper 1: Frameworks for Creating a Common Identifier for a Statewide Data System
https://tinyurl.com/y8sn4smf

Kentucky Center for Statistics
https://kystats.ky.gov/

Minnesota Early Childhood Longitudinal Data System
http://ecIds.mn.gov/

Minnesota Statewide Longitudinal Education Data System
http://sleds.mn.gov/

SLDS Best Practices Brief: P-20W+ Data Governance
https://slds.ed.gov/#communities/pdc/documents/2717

SLDS Issue Brief: Centralized vs. Federated: System Models for P-20W+ System Design
https://slds.ed.gov/#communities/pdc/documents/17542

SLDS Issue Brief: Structuring Data for Cross-Sector Longitudinal Reporting
https://slds.ed.gov/#communities/pdc/documents/16795

SLDS Webinar: Collaboration to Support Data System Modernization
https://slds.ed.gov/#communities/pdc/documents/18373

SLDS Webinar: The Match Rate Dilemma
https://slds.ed.gov/#communities/pdc/documents/8982

SLDS Webinar: Processes for Handling Multiple IDs to Ensure Data Quality
https://slds.ed.gov/#communities/pdc/documents/9697

SLDS Webinar: Use of the Common Education Data Standards (CEDS)
https://slds.ed.gov/#communities/pdc/documents/7906

Washington State Education Research & Data Center
https://erdc.wa.gov/