CALIFORNIA ALL

CALIFORNIA
Data System

Common Identifier Background Paper 1: Frameworks for Creating a Common Identifier for a Statewide Data System

Kathy Bracco and Kathy Booth, WestEd

Introduction

In 2019, California enacted the Cradle-to-Career Data System Act (Act), which called for the establishment of a state longitudinal data system to link existing education, social services, and workforce information.[1] The Act also laid out a long-term vision for putting these data to work to improve education, social, and employment outcomes for all Californians, with a focus on identifying opportunity disparities in these areas.

The legislation articulated the scope of an 18-month planning process for a linked longitudinal data system. The process will be shaped by a workgroup that consists of the partner entities named in the California Cradle-to-Career Data System Act.[2] Suggestions from this workgroup will be used to inform a report to the legislature and shape the state data system designs approved by the Governor's Office. Because the legislation laid out a number of highly technical topics that must be addressed as part of the legislative report, five subcommittees were created that include representatives from the partner entities and other experts. The Common Identifier Subcommittee will

[1] Read the California Cradle-to-Career Data System Act at: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=EDC&division=1.&title=1.&part=7.&chapter=8.5.&article=

[2] The partner entities include the Association of Independent California Colleges and Universities, Bureau for Private Postsecondary Education, California Community Colleges, California Department of Education, California Department of Social Services, California Department of Technology, California Health and Human Services Agency, California School Information Services, California State University, California Student Aid Commission, Commission on Teacher Credentialing, Employment Development Department, Labor and Workforce Development Agency, State Board of Education, and University of California.

help to outline the technical process that will be used to link student records across partner entities.

This brief provides a discussion framework for members of the Common Identifier Subcommittee. It includes background information on the authorizing legislation and a summary of the priorities established by the Workgroup for Phase One of the data system. Next, it weighs the use of an existing identifier to link records on individuals across datasets or using a match process to create a unique identifier. Examples from five states clarify various models for creating a common identifier. Finally, the brief provides an overview of the steps for matching individual data across multiple entities, including some general questions to be addressed at each stage of the process. The brief concludes with framing questions that the subcommittee will consider at its first meeting, in order to recommend which actions should be prioritized in the first phase of state data system development.

The California Cradle-to-Career Data System Act

In 2019, California enacted the Cradle-to-Career Data System Act, which outlined the scope of an 18-month planning process for a state system, allocated $2 million to support that process, and earmarked an initial $10 million toward the development of a state data system.

The Act also laid out a long-term vision for putting data to work to improve outcomes for all Californians, with a focus on identifying disparities in opportunities. By securely linking data that schools, colleges, social service agencies, financial aid providers, and employers already collect, the data system will

- enable users to identify the types of supports that help more students learn, stay in school, prepare for college, graduate, and secure a job;
- provide information that teachers, parents, advisors, and students can use to identify opportunities and make decisions;
- help agencies plan for and improve education, workforce, and health and human services programs; and
- support research to ensure policy effectively supports individuals from birth through career.

Recognizing that the data system will need to be built in phases, the California Cradle-to-Career Data System Act lays out several priorities:

- **Linking existing information in the system.** The first data sets to be linked should be existing K–12 and college data sets, followed by employment and earnings data, early childhood education information, and social services information, although this order can be amended.
- **Guaranteeing privacy and security.** The system cannot be built until clear guidelines and legal agreements have been established to ensure that information will be securely gathered and stored in compliance with federal and state laws and in accordance with privacy best practices, and that the identity of sensitive populations will be protected.
- **Providing information for students, families, and educators.** The system will include an interface for sharing information with teachers, parents, advisors, and students.
- **Facilitating analyses for researchers and policymakers.** The system will link data between agencies to help answer foundational questions about the impact of state policies and investments.
- **Assuring quality.** The legislation addresses the need to improve the quality and reliability of education information, both within and between agencies and other entities providing data.

Priorities for Phase One of the Data System

In the first meeting of the Workgroup, the partner entities recommended that the California Cradle-to-Career data system should be an ecosystem that allows for various tools, processes, and resources to be developed under a governance structure. In its first phase, the state should build a P20W data set that includes early care, K–12, postsecondary, financial aid, and employment information. This data set should be used to create dashboards that provide useful information for practitioners and the public, as well as query tools that allow for more nuanced analyses. The P20W data system should be paired with a clearly defined process to link additional data points as needed to answer inquiries, including requests from outside entities such as researchers, policymakers, and regional partnerships, as well as to foster the secure exchange of information between partner entities. Finally, the Workgroup recommended that the state develop tools that provide information directly to individuals or allow teachers and counselors to better understand the needs of the people they serve. Possible options will be examined at the February 2020 Workgroup meeting, based on tools that have been built in other states or developed in California but not implemented statewide, such as alerting students about the social service and health benefits that they are eligible for, informing educators about services that a student is receiving, or creating an e-transcript service to support college and financial aid applications.

Creating a Common Identifier

As a first step in creating the P20W data set, the Common Identifier Subcommittee is tasked with designing a technical process that will be used to link student records across partner entities in a manner that ensures privacy for individuals. The first key issue for the subcommittee to address is whether to recommend moving forward with an approach that assigns a common identifier by matching individual records using multiple data points, or if the state should adopt an existing identifier that has been assigned by an agency for a specific purpose.

In California, state entities currently use several different methods for linking information (Moore et al., 2017). In some cases, such as matches of education data to the Employment Development Department Unemployment Insurance data set to identify employment and earnings, the student's social security number is used. When a student applies to attend a community college using the common application form CCC Apply, information including first name, last name, date of birth, gender, and high school is sent to the California Department of Education. The department uses a matching algorithm to find the student's State Student Identification (SSID) number in the California Longitudinal Pupil Achievement Data System (CALPADS).[3] The SSID is then shared with the California Community Colleges Chancellor's Office to facilitate matching accuracy for other contexts such as completing mandated federal reporting on career technical education programs. In the absence of a state data system, many California education agencies and institutions purchase information on college-going rates and transfer outcomes from the National Student Clearinghouse, a nonprofit that uses a matching algorithm that references such data elements as first name, last name, social security number, and date of birth.[4]

A 2017 survey of states provides detailed information about the types of data that are included in a state data system, the capacity for linking data across systems, and the specific elements used to link data. States generally use a combination of assigned unique identifiers and element matching processes to link data, with the specific approach dependent on the available data from a given source (Bloom-Weltman,

[3] For more information on SSID see https://www.cde.ca.gov/ds/sp/cl/ssid.asp

[4] For more information see https://www.studentclearinghouse.org/colleges/studenttracker/faqs/

2019). Identifiers such as social security numbers and K–12 identifiers may be included in this algorithm, but they are only one of many variables. Once records are matched, a new common identifier is assigned, which is then linked to all records for that individual (National Center for Education Statistics, 2015).

While the Act requires some postsecondary institutions to track the CALPADS SSID in their data systems, it could be problematic to adopt SSID as the common identifier for the state data system. Using only one data point returns lower match rates than comparing multiple variables (Siddiqi, Sims, & Goff, 2019). Using SSID would mean that the identifier would be missing for students who did not attend K–12 in California and would not be available for children participating in early care that is not part of the K–12 system. Given that employment, social service, and health service agencies do not have an easy way to collect SSID numbers, it could create a significant burden to those organizations to collect and manage K–12 SSIDs and may constrain the number of partner entities that could participate in the state data system. Furthermore, adapting an identifier that was designed for a specific purpose to other contexts may affect the underlying quality of that identifier, particularly if insufficient resources are allocated to training and monitoring other education and service providers to ensure the information they submit meets the data quality protocols used by the originating agency.[5]

Other state identifiers could also be problematic. Children under the age of 16 cannot have a driver's license or state ID card. Undocumented students would not have a valid social security number, and people born outside of the country would not have a birth certificate number or newborn screening number. The lack of this information could inadvertently become a flag about immigration status.

As is noted in the Act, a critical consideration for the development of a common identifier is the need to ensure that data on sensitive populations will be protected. This requirement underscores recommendations from the field. In their Data for the People campaign, the Education Trust-West notes the importance of ensuring that a state data system only be used to help students, never to harm them. Those developing the system must give explicit attention to protecting the identities of undocumented students and families, with clear guidelines for ensuring that a state data system will not put those individuals at risk (Education Trust-West, 2019).

[5] Interview with Amy Fong, December 10, 2019.

Approaches to Creating Common Identifiers

As the subcommittee grapples with questions about how to link data across the different data sources, it may be helpful to consider approaches taken by other states and in California. The examples below are intended to provide a general overview of different ways to approach the process of creating a common identifier.[6]

Washington

Washington Education Research and Data Center[7] builds and maintains a state data system that includes data from early learning, K–12, postsecondary education, and the workforce. Washington uses the vendor Informatica for linking data across the different sectors. Informatica provides a central repository of identifiers (including name, birthdate, and social security number) for every data source, assigns a state identifier to individuals across collections of data, and preserves both merged and unmerged data. The matching process begins with moving source data into a staging database where identifiers are cleaned and standardized, and identity tokens are assigned. Data then moves from the staging database to a Master Data Management (MDM) hub where new data sets are merged with existing information. Each data source has its own matching rules that address differences in the underlying data elements. If a person already exists in the data system, the new information is assigned that person's state identifier. Individuals who are not found in the data system receive a new state identifier. Personally identifiable information, such as names and social security numbers, can be stored securely outside of the data system because data are linked using the state identifier (Sable, 2013).

This approach would work well for providing dashboards and query tools that are based on information in a central warehouse that holds de-identified data. If requests were approved for additional data points on an as-needed basis, those elements could use this same infrastructure to link appropriate records. The Washington model also provides flexibility to develop tools that might require access to identified data for approved individuals. Because identity management and data management are

[6] Members of the subcommittee and representatives from the partner entities will be invited to a professional development opportunity on March 9, 2020, where representatives from other states with well-established match processes will provide more detailed information on their policies and procedures.

[7] Learn more about the Washington data system at https://erdc.wa.gov/

separated in this approach, it offers a potential strategy for protecting sensitive populations.

Kentucky

Kentucky's Center for Statistics[8] maintains the state's data system, which integrates information from the Kentucky Department of Education, the Council on Postsecondary Education, the Education Professional Standards Board, the Kentucky Higher Education Assistance Authority, and the Kentucky Education and Workforce Development Cabinet. Data from each of the participating agencies is submitted through a secure server, where data are validated against a set of criteria from data dictionaries specific to that agency. Newly submitted information then passes through an identity resolution and matching process to link it with records that are already in the data system. The matching process uses an algorithm developed in-house that analyzes multiple data elements to link records for the same individual. The matching algorithm can be customized for specific data sources or batches of data. A scoring process determines whether matches are strong enough to be automatically linked or if they should be reviewed by a staff member. The personally identifiable information used to match records is then removed from the system and replaced with unique system identifiers. Records are stored in a data warehouse that holds only de-identified information, from which reports and analyses can be generated (National Center for Education Statistics, 2014).

As with the Washington model, Kentucky's approach would work well to create dashboards and query tools that are based on information in a de-identified central warehouse, with access to additional data points as needed. Depending on how personally identifiable data are handled after the match, this approach could support data tools that utilize identified data. The homegrown matching mechanism may allow for some additional flexibility and customization as compared to matching tools purchased from a vendor.

Minnesota

Minnesota's state data system[9] is currently in the process of transitioning from a custom-built linking algorithm to a SAS identity management system. The new system is intended

[8] Learn more about the Kentucky data system at https://kystats.ky.gov/

[9] Learn more about the Minnesota data system at http://sleds.mn.gov/

to speed up data loads, incorporate family relationship linking to allow for two-generation outcome reporting, and address data quality issues more quickly.[10] Before making this change, the state conducted a feasibility study that examined eight possible options (seven outside vendors and an internally developed process), the tools and functionality provided by each, and the associated costs. Recommendations from this study ultimately led to the decision to transition systems (Minnesota Department of Education, 2017). In a paper describing the previous custom-built linking mechanism, the staff defined linking as an "ever fluid process," noting that the list of matching rules needed to be evaluated and evolved with each new data set that was added (Minnesota Statewide Longitudinal Education Data System, 2015). In both the earlier version of the system and the new product, a rules-testing process that uses sample data helps to ensure the quality of the matches produced by the proposed rules.

Because this approach is under development, it is difficult to determine whether it could be replicated in California. However, the focus on more sophisticated mapping, such as being able to analyze the multi-generation impact of education, health, and social services, aligns with the goal to improve early care services that hinge on supporting parents and guardians as well as children. In addition, Minnesota could provide helpful insights about problems they are seeking to solve with the redesign.

California

Silicon Valley Regional Data Trust (SVRDT)[11] leverages information from regional, county, and state education, health and human service, and juvenile justice data warehouses to match specific elements such as names, dates of birth, and addresses. Rather than create a unique identifier, an SVRDT internal index records the identifier from each agency that provided services for a specific child. The index only includes information on children who have been served by more than one agency. Because the some of the source data reside in systems that are refreshed on a daily basis, every time new data are integrated or a query is run, the index is updated. The components of the SVRDT match system could inform the linkage process for the California data system because it is designed to bridge the types of data found in education and non-education data sets. This would be useful both in the context of adding elements not found in the P20W data set, as well as for individual- and practitioner-focused tools.

[10] Email communication with Meredith Fergus, January 16, 2020.

[11] Learn more about SVRDT at https://www.svrdt.org/

Children's Data Network (CDN)[12] links information across California Health and Human Services departments and programs using probability match algorithms. The algorithm assigns weights to numerous variables (such as name and address), based on their reliability. It also stores auxiliary information that could inform matches, such as whether children are twins. CDN leverages both machine learning and human reviewers to improve match rates. Once records are matched, the system generates an encrypted linkage key for each individual match. For example, one linkage key is created to connect an individual between Medi-Cal and CalFresh records, and a separate linkage key would be generated for the same individual between Medi-Cal and CalWORKs data. The CDN model could be used for a California data system in two ways. First, CDN provides a model for strong privacy protections that could be replicated with education data. Second, if the tools for individuals and practitioners include health and human service data, they could be built using the existing CDN data set.

Cal-PASS Plus[13] is a clearinghouse for linking K–12 and postsecondary records, which uses an algorithm. Like CDN, the algorithm is enhanced by human review, such as to address collisions—cases where multiple matches are found. Once matching is complete, Cal-PASS Plus removes personally identifiable information and assigns a new identifier. This match process could be used as a model for linking education records for the P20W data set.

California College Guidance Initiative (CCGI),[14] which supports K–12 students in applying to college, matches unique identifiers across systems by connecting those identifiers at the point of application to college. When students begin filling out California Community Colleges or California State University applications from within a CaliforniaColleges.edu account, their SSIDs are linked to their unique application identifiers in the receiving system. The identifier from the receiving system is then returned to CaliforniaColleges.edu so that the accounts are linked in the sending system. Because the match is made using a token that is exchanged between education segments, rather than by trying to compare information about the student, the match rate is 100%.[15] The CCGI process could be leveraged to generate another variable that would improve K–12 and postsecondary matches. CCGI could become

[12] Learn more about CDN at https://www.datanetwork.org/

[13] Learn more about Cal-PASS Plus at https://www.calpassplus.org/Home

[14] Learn more about CCGI at https://foundationccc.org/What-We-Do/Student-Success/California-College-Guidance-Initiative

[15] Personal communication with Tessa Carmen De Roy, January 10, 2020.

one of the data tools for individuals and practitioners if it were scaled statewide to provide e-transcript services for college applications, generate course recommendations for community college math and English courses, and streamline financial aid applications.

Connecting K–12 and Workforce Data

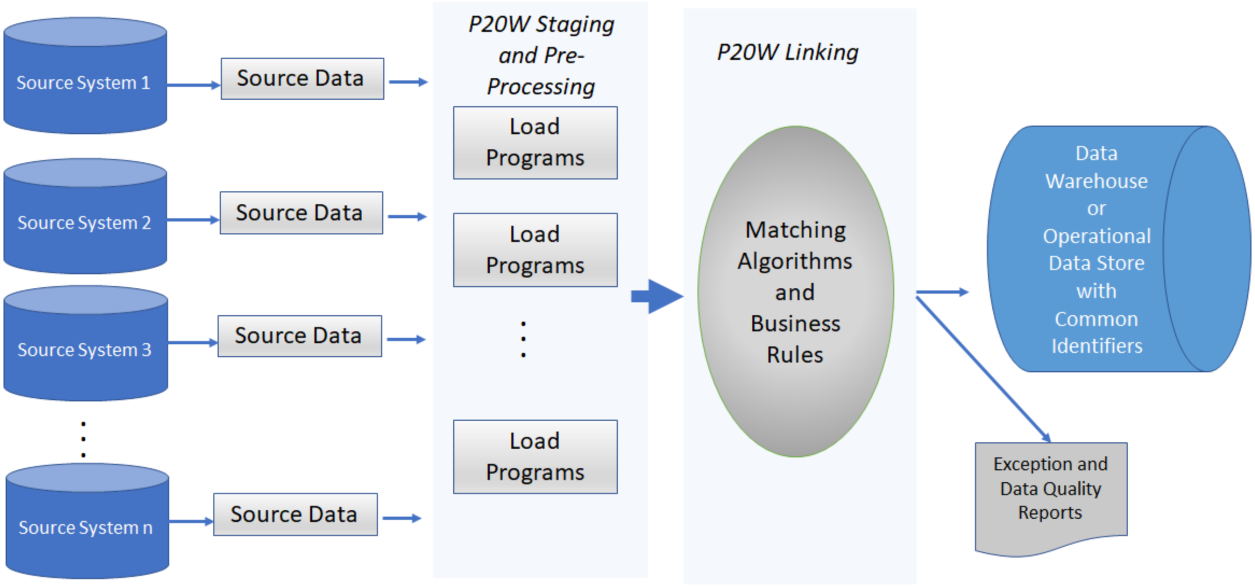
Because K–12 and workforce data systems have little overlap in the personally identifiable information they collect, states have developed novel solutions to create matches, often by incorporating an intermediary data set. For K–12 students who go on to postsecondary education, college information can serve as a bridge because these data sets typically include social security numbers, which are also captured in workforce data sets. But a different bridging data set is needed to link K–12 and workforce data for students who do not enroll in postsecondary education. Many states have turned to the Department of Motor Vehicles as that bridge source.

Idaho uses driver's license and state identifier records, which contain name, date of birth, gender, and social security numbers. To match public K–12 data with workforce data, the State Board of Education attaches a temporary identifier called a labor exchange identifier (LABXID) to student records and sends those records to the Department of Labor. The Department of Labor use a probabilistic matching algorithm to compare the student records to driver's license and state identifier records. Each match is scored according to how closely the names, dates of birth, and genders align between records. Possible matches that do not meet a proscribed threshold are reviewed manually to determine whether they should be linked. Matched data are then assigned a new unique labor identifier (LABUID). Within Department of Labor data sets, the LABUID can be used to identify employment and earnings outcomes by using the associated social security number. The LABUID can also be shared back with the State Board of Education for use in future matches, without requiring the transfer of social security numbers or other personal information between agencies (National Center for Education Statistics, 2014).

Factors in Developing a Common Identifier through a Matching Process

Matching records across data systems requires a multi-step process, illustrated by Figure 1, below. Policies must be developed for each of these stages, to inform specifications for the technical process. If the subcommittee recommends a matching process approach, it can help identify what those policies should be, based on the priorities for the first phase of the state data system.

Figure 1. Common Processes for Linking Data Sets


Source System 1
Source System 2
Source System 3
Source System n
Source Data
Source Data
Source Data
Source Data
P20W Staging and Pre-Processing
Load Programs
Load Programs
Load Programs
P20W Linking
Matching Algorithms and Business Rules
Data Warehouse or Operational Data Store with Common Identifiers
Exception and Data Quality Reports

Pre-processing

Before data can be matched, the underlying data must be pre-processed. Pre-processing can help ensure that data fields are not missing and that the information is correctly formatted, including flagging when there are incorrect values within a list (such as including Asian in a gender data element) or values that are out of range (for example, a birth year of 2050). If the underlying data are not clean, then the matches will not be reliable.

Questions for consideration include:

- Who will be responsible for the pre-processing?

- Which data elements need to be standardized across source systems for matching to occur?
- What needs to be done about missing data?
- Who decides whether individual records or a data set are not of sufficient quality to be included in the matching process?

Matching engine

Matching engine is the technical term for the process of matching records and discerning which ones to link. Many states have opted to use outside vendors such as eScholar, Informatica, and Data Ladder to facilitate the matching process. Others have chosen to develop a homegrown matching algorithm that provides additional flexibility and customizability.

Several big-picture questions for consideration include:

- Will the state use an outside vendor to match records across data sets or will it develop a homegrown product? Could the state leverage a homegrown product developed by another state?
- Given the priorities for phase one, what data elements can be used to link data across different data systems?
- Are there intermediary sources necessary to bridge data across sectors where there is little overlap of data elements (such as K–12 to workforce for those that do not go on to postsecondary)? What bridge sources might be utilized?
- What will be the threshold for determining that two records are matched?
- What should be the process for human review of records that do not meet the match threshold?
- How should testing of the matching rules be implemented?
- How will merging and unmerging records be handled?

Persistence of the identifier and other data

After linking data across different systems, most data systems assign a new identifier specifically for the merged data. However, there are several approaches to the implementation of the new identifier. For example:

- If a new identifier is created, how long should that identifier persist? Should it be used only for a specific inquiry and then destroyed, or maintained for future use and refined with each matching event?

- Should personally identifiable information always be removed once a match is made? If so, should it be destroyed, or should it be maintained separately?
- How could the persistence and location of data help to protect vulnerable populations as well as protect against the ability to identify an individual?
- How does persistence of the identifier impact replicability for research studies?
- How do persistence and location of data impact use cases that require identifiable data?

Refresh schedule

The frequency with which data are refreshed is an important consideration in the matching process. These questions need to be taken into consideration:

- How often is information loaded into the state data system by providers? Does this information vary by data element or data set?
- What are the criteria for updating/refreshing data from providers?
- What are the criteria for adding new data sets/source systems to the state data system?

Quality reports

Once data are matched, a periodic review of the quality of the process must be undertaken. Key questions include:

- What will be the process for reviewing the quality of the matching approach?
- How often should that quality review be undertaken and by whom?
- How and to whom will the results of quality reviews be communicated? How will actions based on the quality metrics be identified?
- How will policies be set related to privacy and security, such as suppression rules?
- If issues arise regarding privacy and security of the data, what actions should be taken?

Preparing for the Subcommittee Meeting

This paper raises a number of questions regarding how best to proceed in developing a common identifier for the California data system, which will form the basis for the February 2020 meeting. Subcommittee members will be asked to weigh in on the following topics:

- Should California create a matching process or adopt an existing common identifier?
- How will vulnerable populations be protected under this model?
- How should this subcommittee structure its work related to developing policies and recommendations?
- What processes should be put in place to ensure that data sets that are not part of phase one implementation can be integrated at a later date?
- Which questions should be directed to other Cradle-to-Career Data System subcommittees, such as the Technology & Security Subcommittee?
- What topics should be covered in a professional development opportunity with representatives from other states that will be held March 9, 2020?

References

Bloom-Weltman, J. (2019). *Statewide longitudinal data systems (SLDS) survey analysis descriptive statistics.* Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2020/2020157.pdf

The Education Trust-West. (2019). *Data for the people: Prioritizing equity in California's state longitudinal data system.* Oakland, CA: The Education Trust-West. Retrieved from https://s3-us-east-2.amazonaws.com/edtrustmain/wp-content/uploads/sites/3/2017/11/01004510/Data-for-the-People-Brief-May-2019-Ed-Trust-West-PDF.pdf

Minnesota Department of Education. (2017). *Feasibility study: Early childhood longitudinal data system (ECLDS*). Minneapolis, MN: Minnesota Department of Education.

Minnesota Statewide Longitudinal Education Data System. (2015). *Person linking for the state of Minnesota's P20W statewide longitudinal data system*. Minneapolis, MN: Minnesota Statewide Longitudinal Education Data System. Retrieved from http://www.ibusiness-solutions.com/wp-content/uploads/2015/08/Person-Linking-for-the-State-of-Minnesotas-P20W-Initiative.pdf

Moore, C., Bracco, K., & Nodine, T. (2017). *California's maze of student information: Educational data systems leave critical questions unanswered.* Sacramento, CA: Education Insights Center. Retrieved from http://edinsightscenter.org/Portals/0/ReportPDFs/Maze-of-Information-Brief.pdf

National Center for Education Statistics. (2015). *SLDS topical webinar summary: Processes for handling multiple IDs to ensure data quality.* Washington DC: National Center for Education Statistics, Institute of Education Sciences. Retrieved from https://slds.grads360.org/services/PDCService.svc/GetPDCDocumentFile?fileId=16363

National Center for Education Statistics. (2014a). *SLDS spotlight: Linking early childhood and K12 data.* Washington, DC: National Center for Education Statistics, Institute of Education Sciences. Retrieved from https://nces.ed.gov/programs/slds/pdf/Linking_Early_Childhood_and_K12_Data_Dec2014.pdf

National Center for Education Statistics. (2014b). *SLDS topical webinar summary: Linking K-12 education data to workforce.* Washington, DC: National Center for Education Statistics, Institute of Education Sciences. Retrieved from https://nces.ed.gov/programs/slds/pdf/Linking_K12_Education_Data_to_Workforce_August2014.pdf

Sable, J. (2013). *Identity matching at ERDC.* Presentation to the Research Coordination Committee.

Siddiqi, J., Sims, P., & Goff, A. (2019). *Connecting the continuum: Longitudinal data systems in North Carolina*. Durham, NC: The Hunt Institute. Retrieved from http://www.hunt-institute.org/wp-content/uploads/2019/06/Hunt-Institute-Connecting-the-Continuum.pdf