Common Identifier Subcommittee Meeting Summary

February 11, 2020

This document provides a summary of the key points that emerged from substantive discussion over the course the day. More information about the meeting, including the background paper and the PowerPoint, are available at https://cadatasystem.wested.org/meeting-information/common-identifier-sub-committee. The website also provides information on the overall process for how the data system will be designed.

The Common Identifier Subcommittee will help to design the technical process that will be used to link student records across partner entities. The February 2020 meeting had the following goals:

- Ground the work of the committee by outlining the recommended scope for phase one of the California data system and key considerations related to creating a common identifier
- Clarify how other states create common identifiers
- Evaluate which models for developing common identifiers best meet requirements for phase one of the California data system
- Develop a recommendation for the process that will be used to establish a common identifier
- Identify priority topics for the March 9 professional development day

The following representatives attended the meeting:

Jonathan Chillas, Association of Independent California Colleges and Universities; Scott Valverde, Bureau for Private Postsecondary Education; Ben Baird, California College Guidance Initiative; Michele Perrault, California Commission on Teacher Credentialing; Todd Hoig, California Community College Chancellor's Office; Pete Cervinka, California Department of Social Services; Glen Miller & Jerry Winkler, California Department of Education; Ben Word, California Department of Technology; Jennifer Schwartz, California Health and Human Services Agency; Amy Fong & Greg Scull, California School Information Services; Sara Pietrowski, California State Board of Education; Jeff Whitney, California State University Chancellor's Office; Joe Hackbarth, California State University, Pomona; Adrian Felix, California Student Aid Commission; Dan Lamoree, Education Results Partnership; Amy Faulkner, Employment Development Department; Eric Goodman, University of California Office of the President; Paco Martorell, University of California, Davis; John Prindle, University of Southern California.

Introductions and Level Setting

The meeting opened with participants introducing themselves and describing how they link records with other agencies. Almost all described using matching processes that rely on a number of variables.

Next, the facilitator provided a description of the benefits of a longitudinal data system, an overview of the California Cradle-to-Career Data System Act, and a description of the process that will be used to craft recommendations for the Governor's Office. Subcommittee participants were encouraged to work closely with their peers on other subcommittees to ensure that workgroup members are able to provide recommendations on behalf of their agencies at monthly meetings on these types of issues.

The facilitator summarized the initial recommendations of the Cradle-to-Career Workgroup regarding the focus of the first phase of data system development, including creating a P20W data set that includes early care, K-12, postsecondary, financial aid, and employment data; making information available via dashboards and query tools; and creating a request process that would allow additional data to be linked for specific purposes. The workgroup will be identifying additional phase one tools for practitioners, students, and families at the February meeting.

Processes for Linking Records

Kathy Gosa, a representative of the federal State Longitudinal Data System Support Team, provided information on how records are typically matched for a state data system and outlined some of the key issues that need to be addressed when creating a common identifier.

Next, she walked through the challenges of using a single existing identifier, such as a social security number or a K-12 identifier like the California Department of Education state student identifier (SSID). Other states have not adopted a single existing identifier because each data partner typically collects slightly different information. By creating a set of business rules that determine common variables between each data set, the state can link records without requiring each agency to change its local process for assigning unique identifiers. This is particularly important to ensure that identifier data is of high quality—most state agencies devote significant effort to ensuring that their identifiers are assigned appropriately and have mechanisms in place to address cases where multiple individuals have been assigned the same number or an individual has been assigned more than one number.

The subcommittee discussed whether California should adopt a single existing identifier or use a match process that links multiple data elements. Most subcommittee participants felt strongly that it would not be appropriate to use SSID as the state's common identifier. Instead, they recommended that the state data system use a match process that pairs the identifiers used by individual partner entities (like SSID) with additional elements such as name, date of birth, and address. Several subcommittee participants described how they have been able to successfully match records using multiple variables and reported that matching a range of variables across partners resulted in stronger overall match rates because there were more data points to distinguish similar individuals. These participants also noted that individual identifiers such as social security numbers and SSID are important components of those match processes, particularly in cases where multiple possible matches are returned.

One participant noted that while a multi-variable match process would work well to link records, it may be helpful to extend some existing identifiers in contexts where none are currently in place. For example, it might be possible to extend SSID in the context of early care or use social security numbers for private colleges.

The Employment Development Department representative noted that the department only conducts matches using social security numbers. They would not be able to adopt SSID and will have to ensure they can legally integrate employment and earnings data into a matching algorithm.

Some subcommittee participants noted that a match process could be used to create a new identifier that can then be used for the state data system. Finally, one participant indicated that the most important thing is to create a well-thought out process that can be used to ensure the consistency and validity of data.

Data Linking Models from Other States and California

To provide examples of how California could implement a matching algorithm, Kathy Gosa described the processes used in Kentucky, Minnesota, and Washington. All three states will be coming to California on March 9, 2020 to meet with the Common Identifier Subcommittee and the Technology & Security Subcommittee in an optional professional development day. Subcommittee participants identified questions they would like the states to address, including:

- What was the impetus for creating the data system?
- What were the costs of building and maintaining the data system? What processes did the state use to develop the system? How long did it take to create the system?
- Does the data system live on physical servers or does the state use a cloud provider?
- Is personally identifiable information stored separately from other data? If they are separated how are data linked between the two?
- To ensure that records cannot be reidentified, what policies has the state set regarding the minimum numbers of students included in data reports/displays and how complementary suppression should be applied?
- Why did the state choose to either build or buy a matching solution? If the state made a change to its matching solution, why?
- When the match returns multiple possible individuals, how much time is spent to resolve each collision or duplication? What are the business processes used to handle exceptions? How much of this process is automated and how many individuals support the computerized processes?
- When there are multiple people assigned to the same identifier or the same individual has been assigned multiple identifiers, does the state make changes to underlying records? If so, how does the state track those changes? Are the changes made only in the state data system or do the partner entities change their records? How far back in the individual's record are changes made? How do these changes affect what is shown in dashboards or produced in reports?
- How does the state data system handle incorrect identifiers submitted by the partner entities?
- What types of support does the state provide to partner entities regarding the state data system?

Next several California-based projects that have linked data shared their approaches, including Children's Data Network, Cal-PASS Plus, and California Colleges Guidance Initiative. The facilitator also provided a summary of the process used by the Silicon Valley Regional Data Trust. These projects have successfully address priorities identified by the Cradle-to-Career Workgroup including working with real-time data and integrating health and human service records within California's legal constraints. Most notably, most of these projects have not created a common identifier but use strategies such as indexes, linkage keys, and token systems to link records. Subcommittee participants asked a number of clarifying questions to better understand these approaches and discus lessons learned about matching processes.

Key Decision Points in Creating a Common Identifier

Kathy Gosa outlined five critical aspects that states must address when creating a common identifier: pre-processing, the matching process, persistence of the identifier and other data, refresh schedules, and quality reports. The subcommittee then discussed the best way to develop recommendations on these topics. They determined that they will develop preliminary recommendations at the March subcommittee meeting, after the workgroup has set the direction on the remaining components for phase one of the data system and subcommittee participants have had a chance to learn from other states at the professional development day. During April, the subcommittee participants will work with other experts from their agencies to refine the proposal. The subcommittee determined that they should hold an additional meeting in May to revise their proposal based on this input.

The group spent the remainder of the meeting conducting an initial review of the five areas.

Pre-processing. To get a sense of the data elements that would need to be standardized across sources, the subcommittee listed the elements that they currently use or could leverage for matching purposes. This list will be refined via a Google document before the March meeting, including ranking whether each data element would be considered reliable for matching.

Matching process. Kathy Gosa described how states have filled in missing elements necessary to conduct matches, such as collecting driver's license numbers from students and then sharing the driver's license numbers with the Department of Motor Vehicles to secure associated social security numbers, so that K-12 records can be matched to employment data. Subcommittee participants noted that it would be helpful to invite the Department of Motor Vehicles to participate in future meetings to determine whether a similar strategy should be pursued. Other participants noted that licensure and apprenticeship data might also serve this bridging role.

Persistence of the identifier and other data. The group discussed the concept of temporary versus enduring identifiers. Many states create a common identifier when they match records that is used consistently within the state data set, while others create a unique identifier at the time that they conduct a match for a specific purpose and then destroy that number. Temporary identifiers allow for stronger privacy protections but make it difficult to replicate analyses. Temporary identifiers can also make it harder to improve upon matches as more data elements are gathered across data sets and create a more time-intensive process because the validity of the match must be re-examined each time.

When considering whether to have fixed or temporary identifiers, subcommittee participants noted that the answer may be driven by how strong the match needs to be. For research purposes, it can be acceptable to have a less rigorous match, whereas if the data are being used to provide services to specific individuals, the match has to be more precise. Subcommittee participants noted that California might need to have two different processes—and associated thresholds for matches—for the P20W data set and research requests, versus for practitioner- and student-focused tools. The discussion at the next meeting should examine these factors in the context of the types of tools that the workgroup prioritizes.

The California Commission on Teacher Credentialing representative described how the commission has a unique identifier that is used within their data warehouse. However, when researchers request information from the data warehouse, new unique identifiers are created that are specific to that requestor. In this way, a researcher can request multiple queries and get consistent information. However, because each requestor has a different set of unique identifiers, there are greater privacy protections. The state data system could deploy a similar hybrid approach.

Others noted that the decision about permanent versus temporary identifiers would also be influenced by how federated or centralized the system is, which will be discussed at the Technology & Security Subcommittee meeting later this month, and by how much information is stored in the P20W system.

Finally, another participant noted that it would be important to discuss the feasibility of permanent and temporary identifiers with the Legal Subcommittee and to clarify the benefits and drawback of each approach in the context of the goals for the state data system.

Given the complexity of this topic, the group recommended that it be addressed at both the March and May 2020 meetings.

Refresh schedules. Because this decision will require clarity about the types of tools that will be developed, the subcommittee suggested waiting until May to discuss this topic, after the workgroup has finalized its recommendations for phase one tools.

Quality reports. The subcommittee recommended addressing this topic in both March and May.

Next Steps

The subcommittee requested that Kathy Gosa provide several resources to support the design of the California system, including:

- a summary of matching products that are commonly used by other states
- examples of the rules used by other states for matching, particularly related to thresholds, human review processes, testing procedures, and handling of merges and splits
- a summary of the benefits and trade-offs of temporary versus permanent identifiers

In addition, they suggested that information be gathered on the match process used by the Migrant Student Information Network.

Finally, the subcommittee added an additional meeting on May 7, 2020, where the discussion of the five critical aspects of creating a common identifier will be continued.