Common Identifier Subcommittee Meeting Summary

March 10, 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 background reading and the PowerPoint, are available at https://cadatasystem.wested.org/meeting-information/common-identifier-sub-committee.

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

- Ground the work of this committee by providing an update on the recommended scope for phase one of the California data system
- Develop a process for establishing pre-processing standards
- Recommend the persistence of the identifier
- Identify standards for matching
- List considerations for quality reports

The following representatives attended the meeting:

Jonathan Chillas, Association of Independent California Colleges and Universities; Joanna Murray, 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; Jayson Hunt & Akhtar Khan, California Department of Social Services; Jerry Winkler, California Department of Education; Janet Buehler, California Department of Technology; Jennifer Schwartz & Chris Krawczyk, California Health and Human Services Agency; Amy Fong & Greg Scull, California School Information Services; Jeff Whitney, California State University Chancellor's Office; Patrick Perry, 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.

Workgroup Update

The meeting opened with the facilitator providing an update on the topics covered in other subcommittee meetings during the month of February and decisions made by the California Cradle-to-Career Workgroup on their February 26 meeting.

Pre-Processing Standards

Note: Pre-processing is the work done to prepare data elements for matching.

Kathy Gosa, a representative from the federal State Longitudinal Data System Support Team who was helping to facilitate the meeting, handed out a chart of possible data elements that could be used for matching purposes, based on input from each partner entity. The grid included color coding to indicate the degree of reliability for a match process. Subcommittee participants provided further input and clarifications on the grid, including:

- The early care representative from the California Department of Education (CDE) noted that elements had been marked in yellow because they would not be reliable for matching when taken individually, but when combined there was sufficient information for a match.
- The Bureau of Private Postsecondary Education representative indicated that, while legislation only requires that they capture social security numbers for students, colleges are collecting additional information that could be reported to the state including high school (a strong element), as well as name, address, phone number, and email address (weaker elements).

One participant noted that data quality can vary in a number of ways. For example, the state agency could have clear standards, but individual institutions might report faulty data. Also, data points may be of higher quality at different points in a student's experience, such as having more reliable information for students who enrolled compared to those who applied but did not enroll. Kathy Gosa clarified that states normally do an extensive vetting process about each match element when data sets are brought into longitudinal systems to document this level of nuance.

Several subcommittee participants who had experience with matching noted that while preliminary rules can be established based on logic, the algorithm must be refined using the actual data. Furthermore, processes must be put in place to revise the rules over time.

The group also weighed in on how much pre-processing should be done by partner entities versus by the entity that hosts the state data system. Participants argued that partner entities should have primary responsibility for pre-processing given the scale of the data system and the need to have a sophisticated understanding of many different data sets. The effort to develop the algorithm, including validating the matches, will be significant in the development phase, but should be less intensive over time, particularly if California uses a master index approach like the one in Kentucky. It will be important for the host entity to be in regular communication with partner entities to integrate new elements or make adjustments based on changes to existing elements.

The subcommittee noted that the decision about pre-processing has implications for staffing. Sufficient resources will need to be allocated to each partner entity to support the work of developing and tuning the algorithm. Appropriate staffing is directly correlated to data quality—if resources are not devoted to ensuring that underlying data and match algorithms are correct, then the state data system will return misleading or erroneous results. Furthermore, for use cases like linking health and human services data to determine whether students are receiving services, matches may need to be done with local providers in order to provide real-time information, which would mean providing support and oversight to thousands of entities.

In considering how the file format could be aligned across partner entities, subcommittee participants recommended that general constraints should be established at the outset, based on what is feasible for the partner entities. For example, the schedule for uploads should be predicated on existing data flows, rather than requiring partner entities to adjust reporting requirements from their constituencies. However, once this schedule is established, partner entities would be expected to report within those timelines.

The subcommittee also recommended that partner entities be allowed to provide information in its native format, rather than requiring everyone to adopt a common standard in their data sets, particularly given the fact that many partner entities are working with aging legacy systems that cannot be easily reconfigured. The host entity should be tasked with converting the data sets into a common file format.

One participant noted that it will be important to evaluate how much work the host entity should do to improve data quality. For example, if high standards are set at the point that information is loaded into the system, such as rejecting a file because elements are missing from a handful of records, technical assistance will be needed for both the partner entities and the individual schools or providers that generate information for state agencies.

This balance of time versus quality would also need to inform decisions about whether the host entity would be empowered to adjust data to maximize matching or if the partner entities would be expected to work with data providers to fix data in local systems when inconsistencies are found—for example, changing a student's racial designation to align K-12 and college records. Discrepancies in data can lead to problems when similar information is displayed by different parties—for example if the state data system produces college readiness reports that differ from CDE's School Dashboard due to adjustments to student characteristics that were implemented as part of the match process. These types of discrepancies can fuel distrust.

Subcommittee participants noted that data quality thresholds can vary based on the use case—for electronic transcripts, it is vital for information to be as accurate as possible. When funding is based on state data, accuracy becomes more urgent for data providers. For research, a lower standard of match quality may be acceptable. One approach for the state data system would be to clearly flag the level of data quality, both to explain discrepancies in results and to determine whether data are of sufficient quality for their intended use. But given that data matches will shape all other information provided by the state data system, the participants recommend that the highest quality possible data are needed for the match algorithm.

Persistence of the Identifier

Note: Once data are matched, states generally assign a state identifier to the matched record. However, they differ regarding whether that state identifier remains permanently attached to the matched record or if new identifiers are created either periodically or at the time of each data request.

After breaking into small groups to discuss the persistence of the identifier in the three use cases (P20W data set, data request process, tools for practitioners and individuals), subcommittee participants voted unanimously to implement a permanent identifier within the state data system. Key aspects of their discussion on this topic included:

- A permanent identifier allows the validity of a match to improve over time, particularly as new data sources are integrated. It would be valuable to adopt an index approach that tracks information on changes to match status—such as cases where previously matched records are found to be different people.
- States like Minnesota that regenerate the state identifier every six months have weaker legal liability for data breaches. California may not need that level of caution, because those receiving data have stronger incentives to keep it secure, although the state data system should have a protocol in place to create new identifiers if the data set is breached.
- California should adopt the common practice of creating new unique identifiers for data sets that are shared for authorized research purposes. A crosswalk of the research identifier to the

state identifier should be kept, so that consistent data could be pulled at a future point to extend an analysis. Use-specific identifiers could also potentially be used in other contexts, such as for authorized tools that access state data to support service delivery between education and social services.

- Generally, the state identifier should only be used within the state data system and not shared. For example, if the state identifier were used to generate an electronic transcript, the state identifier should not appear on the transcript. However, there may be circumstances where it is appropriate to release an identifier. The governance structure should address how to make decisions about releasing state identifiers. In the governance discussions, attention should also be paid to the legal obligations of the host entity or vendors related to managing and releasing identifiers.
- The question of whether the state identifier is stored with personally identifiable information or in a crosswalk table should be addressed by the Technology & Security Subcommittee.

Creating Standards for Matching

Note: The standards used for matching are highly complex and directly related to the software and hardware associated with the state data system. Therefore, the discussion focused on how to identify technical solutions for implementing a match process.

Using a feasibility study and request for proposals from Minnesota as an example, the subcommittee participants paired up discuss how to evaluate potential technical solutions for the match process and what types of information should be gathered. Then, subcommittee participants were asked to vote on whether to conduct a feasibility study, release a Request for Information (RFI), or do both.

Most subcommittee participants voted to go straight to an RFI, with several key considerations:

- Information should be gathered from state agencies that already have a match process, to see if any of these existing tools could be scaled.
- There is lack of clarity about what would be included in the RFI because the scope of the state data system has not been finalized. Developing an RFI may be helpful in the design process because it will force some decisions about what the data system must be able to do. It can also help develop a ballpark for costs.
- It will be vital to have the partner entities weigh in on the content of the RFI and ensure that it will produce information that is helpful to the design process.

One participant noted that the California Office of Statewide Health Planning and Development recently sent out an RFI that could be referenced in developing a document for the state data system.

Amy Fong and Greg Scull from the California School Information Services, Eric Goodman from the University of California, Patrick Perry from the California Student Aid Commission, Chris Krawczyk from California Health and Human Services Agency, and Jerry Winkler from CDE volunteered to participate in a process with the California Department of Technology to develop an RFI.

In addition, one participant suggested that it would be helpful to do a readiness study to better understand the status of elements that would be used for matching at each partner entity.

Quality Reports

Note: Quality reports help to flag potential issues with the underlying data used in the match process and to tune the match algorithm.

After hearing about some examples from other states and California's migrant education system, subcommittee participants revisited concepts discussed earlier in the meeting, including the value of providing a score for each match to indicate the level of validity, which could inform whether and how it is used.

Participants also noted several concepts that could be included in the RFI related to the question of quality reports, including:

- Whether a cluster approach is used in matching
- How data quality fluctuations for the same record are addressed
- How the match algorithm uses reliability of elements over time to improve its formula

Finally, one participant suggested crowdsourcing the best possible match methodology by putting out a call to academia and vendors to develop an optimal solution that would be available to everyone in the public domain.