Technology & Security Subcommittee Meeting Summary

February 18, 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/techsecurity-sub-committee. The website also provides information on the overall process for how the data system will be designed.

The Technology & Security Subcommittee will develop technology specification requirements to address data structures and privacy considerations. The February 2020 meeting had the following goals:

- Ground the work of this committee by outlining the recommended scope for phase one of the California data system and key considerations related to the underlying data structure
- Clarify how other states and California-based data systems structure their data sets
- Evaluate which data structures best meet requirements for phase one of the California data system, including identifying constraints and opportunities that are unique to California
- Identify priority topics for the March 9 professional development day

The following representatives attended the meeting:

Formeka Dent, Antelope Valley Union High School District; Helen Norris, Association of Independent California Colleges and Universities; Jason Piccione, Bureau for Private Postsecondary Education; Ben Baird, California Colleges Guidance Initiative; Amarjot Biring, California Commission on Teacher Credentialing; Barney Gomez, California Community College Chancellor's Office; Rodney Okamoto & Alan Nakahara, California Department of Education; Karissa Vidamo, California Department of Social Services; Vitaliy Panych & Janet Buehler, California Department of Technology; Dan Lamoree, Education Results Partnership; Deanne Wertin, California Health and Human Services Agency; Greg Scull, California School Information Services; Ed Hudson, California State University Chancellor's Office; Gurinder Bains, California Student Aid Commission; Noah Bookman, CORE Districts; Ruby Raines, Employment Development Department; Jenni Abbott, Modesto Junior College; Steve Ambrosini, Richard Gold & Marcy Lauck, Silicon Valley Regional Data Trust; Matthew Linzer & Hooman Pejman, University of California Office of the President; and Douglas Leone, Labor and Workforce Development Agency.

Introductions and Level Setting

The meeting opened with 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.

Participants introduced themselves and described the types of data their agencies or organizations collect and how they display or link that information.

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.

Identifying a Possible Data Structure for the First Phase of Development

Baron Rodriguez, a national expert on state data systems, described the pros and cons of federated and centralized data models. The subcommittee discussed their experiences with each approach and reflected on data structures that would support a P20W data set and a research request process. One participant noted that the state system is likely to evolve over time. It may begin as a federated approach and later evolve into a centralized approach, once trust is built between partner entities and the public and when issues of how to match and display information are determined. Participants stressed the importance of clarifying to the public how their data will be used, including ways it will be support the common good and not just benefit specific agencies.

One participant indicated that for use cases like dashboards and query tools, it is helpful to have centralized data sets that can quickly return information, rather than pulling data from disparate systems and building results from scratch for each request. When one participant noted that there are legal restrictions that make it difficult for some agencies to house data outside of their systems—particularly given regulations that govern the sharing of health, financial aid, and employment data—another noted that different data models may be needed for different use cases.

The subcommittee discussed the importance of tools for practitioners and individuals, which frequently require identified, real-time data in order to be actionable. One participant wondered if it would be better to have regional collaboratives build real-time data systems with identified records and focus the P20W data system and request process on agency-level, de-identified, historical information.

Participants noted that it would be valuable to create a data catalog and a library of research and metric methodologies so that various external entities could conduct research in a consistent fashion. Several members raised questions about the governance process for data requests, such as who would determine whether requests are granted, how research methodologies would be vetted, and who would review results to ensure that reasonable conclusions have been drawn. They noted that it will be vital to ensure there is adequate staffing and funding to ensure knowledgeable parties are part of this review and validation process. One participant suggested the entities requesting data should be charged to help cover the costs of reviewing requests and subsequent results, as well as to increase a sense of ownership from those requesting information.

Other concerns that were raised included the risks of putting comprehensive information on individuals all in one place; whether real-time data is important for dashboards, query tools, and research requests that use de-identified data; whether the information would be stored in the cloud or on servers; and whether historical information in the state data system would be preserved longer than it is kept in individual agency repositories.

Several of the education partner entities, including the California Department of Education, California Community Colleges, California State University, University of California, and the Association of Independent California Colleges and Universities presented a high-level draft concept paper about data structures (the full document is posted with the meeting materials). They proposed that the data system should:

- Use a hybrid approach, rather than picking a federated or centralized model
- Ensure that partner agencies maintain control over their own data while creating responsibility to share specific data elements
- Create the highest levels of privacy and security
- Store data in the cloud
- Be vendor agnostic and use open-source technologies
- Produce curated data sets that provide a single source of truth
- Foster transparency including clear and consistent data definitions, a data catalog, and notifications of data changes

The group discussed the proposal at a high level, including exploring the role that data lakes, which some partner entities are developing, could play in relationship to the state data system.

Next, Baron Rodriguez described specific federated and centralized data models that have been implemented in Virginia, Nevada, Kentucky, Minnesota, and Washington, followed by brief descriptions from subcommittee participants of data models that have been implemented in California by Cal-Pass Plus, the Children's Data Network, California Colleges Guidance Initiative, and the Silicon Valley Data Trust. Participants had an opportunity to ask clarifying questions to better understand these approaches.

Optimal Data Structure for the P20W Data Set and the Research Request Process

Participants self-selected into one of two groups—one focused on building a P20W data set and one focused on providing data for research requests. Each was tasked with designing an optimal data model that addresses factors such as the level of effort required to align data, the ability to view historical data, data recency, security controls, and governance complexity.

In the report outs from these discussions, both groups indicated that a hybrid approach would be most appropriate. They noted that, generally, a centralized approach is preferable because it is easier to create consistent governance rules, ensure data is of higher quality, provide faster access to information, and manage the workload on partner entities. Batch uploads to this system could be timed to take into account refresh schedules managed by the partner entities. However, in cases where laws prohibit data from being stored outside of agency systems, federated aspects should be incorporated. Federated structures may also prove to be preferable for the student- and practitioner-facing tools that the workgroup will prioritize later this month. While there was a general consensus on this hybrid approach, one participant expressed concerns about cost and wondered if it would be more expensive to build a centralized system than a federated one. Another felt it is important to treat personally identifiable information in a centralized system differently than other data points.

Both groups indicated that it would be preferable to have a subset of information reside in the P20W data set, rather than compiling all possible data points available from partner entities. Holding all data

creates security risks that are much harder to manage. The subcommittee concurred with the workgroup recommendation that contents of the P20W data set should be determined based on the state's goals for the system, adding that partner entities should have control over which data points they elect to share.

Some subcommittee participants noted that designing a system in the abstract is challenging. The current level of specificity (a broad vision of improving education and health outcomes for Californians, using dashboards, query tools, and research request processes to help a variety of audiences make decisions) is not sufficient for designing the data architecture. Having much more specific direction from the Cradle-to-Career Workgroup would enable this subcommittee to be more effective in creating technical specifications. For example, it would be helpful to understand who would use the dashboards and what they would do with the information provided. One participant noted that the question of which data elements to link must be made before the data system is built—not afterwards. Another stated that it will be important for the workgroup to help set expectations with the research community about the amount of information that can be realistically be made available.

Next Steps

Subcommittee participants outlined the following questions for the Legal Subcommittee:

- Will the state data system need to track individual permissions for using data?
- Can entities that conduct analyses using state data system information be required to share their results? How do those requirements vary based on the funding source for the request?
- Can historical, de-identified data be stored indefinitely, or would they need to be destroyed at the same time that source agencies delete older files?
- What are the legal rules around data disclosure? For example, if data are shared with an outside entity through the request process, could someone use a Public Records Act request to require that outside entity to provide unitary data?
- Do any of the partner entities have legal restrictions that prevent them from putting data into a centralized system?
- How can the partner entities hold others accountable for the data they receive? How does this relate to indemnity policies?

They also identified the following questions for other states:

- How have other states dealt with security requirements related to financial aid?
- How have they linked education and health information?
- Where is their data physically being hosted (cloud/server)?
- Are research designs and reports shared back when outside entities make a research request?
- What level of effort is needed for each request made in a federated system?
- Are there differences in performance for federated versus centralized systems?