A Roadmap for Linking Administrative Data in California

CALIFORNIA POLICY LAB

A long-run vision for state data

California is the home of Silicon Valley but the state's data systems lag surprisingly far behind. Forty-four states systematically link student-level postsecondary data to K-12 or workforce data, but California does not.[i]

Recent calls to create a "P-20W" or "cradle-to-career" longitudinal data system in California have sparked interest in better connecting the state's administrative data. These efforts are heartening, given the value that such data could provide state policymakers. **But a narrow focus on educational data risks missing the bigger picture.** Education interacts with other domains, and student success is bolstered by healthy families, safe neighborhoods, and economic well-being. The state needs a long-term vision for making better use of *all* of the state's data for improving policy outcomes. Several recent reports by expert panels emphasize the critical role that linked data could play in other areas, like anti-poverty programs and criminal justice.[ii]

California is the fifth largest economy in the world. It needs data systems that help the state understand its residents and serve them better. We propose here a roadmap toward that goal: (1) **create a new, independent office** with the mandate and expertise to link data across siloes, (2) **sequence the linkage process** by starting with education and expanding outward, and (3) **establish streamlined governance** that makes data available to improve state policies and programs.

The case for linked data

Most complex policy problems cut across sectors. Homelessness, for example, intersects housing, health, and criminal justice. And educational achievement is driven not only by teacher and school quality, but also family well-being, neighborhood factors, and student health.

California does not have the tools necessary to address these complex problems because its data is siloed in ways that obscure good information about potential causes and solutions. **We cannot solve a Rubik's Cube when we see only one side.**

We currently cannot answer simple questions such as:

- Which California high schoolers are prepared for college?
- How many families are claiming the safety net benefits for which they're eligible, and what consequences does this have for their children's life outcomes?
- Which programs successfully help formerly incarcerated individuals reintegrate into society?

Linking California's data assets could help answer these questions and lift tens of thousands out of poverty, improve our schools, and save the state tens of millions of dollars through program improvements. These data could be used by researchers to inform policy, by households to inform school and neighborhood choice, and by policymakers to inform program improvements.

A state data linkage hub: Statistics California

California needs a centralized authority for linking the state's administrative data. Legislators are focusing on new *datasets* and *data systems*, which is a step in the right direction.[iii] **But what the state truly needs is a new *office* with a clear mandate to link the state's core data assets, a clear set of tools for doing so, and governance that ensures data are used to inform program improvement.** Think of it as the state's Census Bureau – or "Statistics California."

To succeed, this data linkage hub needs all five vowels -- AEIOU:

- the statutory **Authority** to compel data from contributing data providers
- the technical **Expertise** to link, protect, curate, and provision large and highly sensitive data
- the structural **Independence** to avoid undue influence by any particular data contributor
- an **Open source orientation** so the office can easily build and adapt solutions over time to avoid wasting millions on proprietary technology procurements
- an **Unwavering commitment** to privacy, security, and confidentiality

Similar linkage hubs have generated substantial benefits for other states. With a modest investment, a state like California could reap large returns through improved program performance and better outcomes for the people the programs serve.

How would an independent data hub work?

There are several versions of how a data linkage hub could work. **We describe here one version of how such an office could function, drawing on best practices from other states.**

An authorizing statute would establish the hub and designate a timeline on which certain agencies would be required to contribute data. Statistics California would work with each data provider to construct a pipeline for regularly transferring, transforming, matching, and loading data into the Statistics California data hub. Each agency's data would be updated at least annually, and sometimes more frequently. This process would be automated to the extent possible, to minimize ongoing costs and to enable more frequent updates over time. Contributing data providers would keep their original data, and continue to use it for regular operational purposes.

Statistics California would maintain the confidentiality and security of the data and would store data in linked or linkable formats. Linkage hub staff would curate the data and create data catalogs, dashboards, and common data products to match user needs. The hub would also extract custom datasets for approved projects (governance described below), and would make those data available in secure virtual enclaves, available only for specific purposes and users. Some data uses might be approved on an ongoing basis for which user-restricted APIs could be established. Results would be reviewed to ensure privacy protections.

Sequencing for Success

The ideal linked data system would combine education, workforce, criminal justice, safety net, health, and other data. Experience from other states suggests that education and workforce data are the best place to begin, with incremental linking of datasets from other areas. The relevant state data systems are already in readily linkable formats, and indeed the relevant agencies have prior experience with most of the necessary linkages.

We propose that Statistics California pursue a sequence of data linkages starting with higher education, workforce, and K-12, and then expanding over time to include health and human services, other backbone data, criminal justice records, and early childhood. This sequence is based on our extensive experience with state data across several domains, and our understanding of the data itself, its reliability, and its usefulness for informing policy and improving lives. Though not exhaustive, this roadmap would cover the next several years.

Education and workforce

CALPADS, the higher education segments, and EDD's quarterly earnings records can all be easily linked and will yield important insights about educational progression and career readiness. Student Aid Commission data on FAFSA and Cal Grants would be another valuable, and simple, addition. With cooperation from the data contributors, a working prototype for these linked data could be created in under a year.

Start with the higher education segments. CCC, CSU, and UC all maintain student-level data on demographics, enrollment, course-taking, and performance. These systems share similar identifiers and can be easily linked and mapped to a similar data model.

Link workforce next. EDD has person-level quarterly earnings records that can easily be linked to higher education records. Many research use cases will require income and employment outcomes, which are often a measure of program success.

Then merge CALPADS. Merging CALPADS will require "probabilistic matching" to retrospectively link students. Establishing a unique student identifier upon school entry would help solve this problem going forward.

Health and Human Services

Over the past two years, the Health and Human Services Agency has undertaken an ambitious project to link data among the many departments it oversees. It is currently building a CHHS Research and Data Hub that will bring together birth and death records, Medicaid claims and hospital discharge data, and social safety net program participation, among others. We anticipate CHHS to proceed in parallel with education-related efforts. All that is required is for both efforts to coordinate and ensure that each system is designed in a fashion that allows for the CHHS data to be linked into the state linkage hub. That step is not overly burdensome, but omitting it would miss a prime opportunity to link to existing efforts.

Backbone data

DMV records and FTB tax records are a logical next step because they fill out the population frame, covering most adult residents, and also providing reliable and updated residency and household composition information. These data are key to understanding mobility and household-level effects, and the tax records also give insight into important government benefits (e.g., EITC, CTC), homeownership, and health coverage. State law changes could enable better use of such data for research purposes.

Criminal justice

California has over 500,000 residents currently incarcerated or under some form of supervision, and a much larger number have a history of justice involvement. With "realignment" and Props 47 and 57, we have undergone one of the nation's most dramatic criminal justice reform efforts in recent history. Criminal history data from CADOJ, and prison and parole data from CDCR, could yield valuable insights about violence prevention programs, links between crime and other social indicators, and programs that aid successful reintegration following incarceration.

Early childhood

California currently has no statewide early childhood data system. For that reason, the state should focus on creating data standards and reporting requirements for such a system, and in doing so, should design it so that it can later be seamlessly integrated with other education data.

Future linkages

A successful linkage infrastructure would eventually allow for other linkages, for example to housing and transportation information, to spatial data on environmental conditions, and to valuable data assets at the local, regional, or federal levels, all while maintaining individual privacy and applying strong governance procedures.

Depoliticizing data governance[iv]

Most other states are ahead of California in linking state administrative data, and we can learn from their mistakes. Their experience teaches that the main risk is a data system that is built but never used. This often arises from failures of governance: **Where states have politicized their governance process, and given veto power to each data contributor, the outcome is inertia.**

Successful governance processes empower subject-matter experts to review project proposals based on well-defined technical criteria, and include provisions – such as defaulting to approval if a proposal is not affirmatively rejected within a designated timeframe – to ensure that the approval pipeline does not become a project graveyard. Another aspect of successful processes is that they separate concerns over privacy and data security from those about the substantive content.

For each use under consideration, we recommend an initial legal and privacy review be conducted by data hub staff. They would verify that the project was legally permissible and would not jeopardize individual privacy. They would be charged with "getting to yes" by suggesting alterations that would alleviate any concerns, including by masking identifiers, inserting statistical noise, or using secure multi-party computing.

Following legal and privacy review, we recommend that each proposed use be reviewed by a Data Use Approval Council. Such uses might be for a specific project (e.g., to evaluate a new home-visiting program) or for ongoing use by a particular user type (e.g., to regularly generate employment outcomes for each college campus by major). The Legislature could add and prohibit specific uses by statute. The Council would consist of subject-matters experts from the agencies whose data is implicated in that particular use. They would publicly report their decisions using a set of criteria determined by the Legislature. These criteria should focus on furthering the public interest and on good project design, and should expressly disallow interests that specific data contributors may have in avoiding embarrassing results. As with the recently enacted federal Foundations for Evidence-Based Policy Act, if the proposed use would generate new evidence, there should be a presumption of approving the project absent a compelling reason not to.[v]

Our expertise

The California Policy Lab's mission is to improve people's lives by generating evidence that transforms public policy. We were launched at UC Berkeley and UCLA in 2017 with the express purpose of helping California's governments better link[vi] and analyze the data they collect. In two short years, **we have established a cross-sector secure data linkage infrastructure** with data from over a dozen state and local agencies, including data on health, taxes, criminal justice, and education. With special attention to maintaining individual privacy, we are bridging data siloes with research that cuts across policy domains.

Together with state agency partners, we are using linked data to encourage more eligible households to claim the Earned Income Tax Credit and the state's Cal Grant college scholarship. We are also using linked data to improve the state's safety net programs and to reduce homelessness and crime in Los Angeles, San Francisco, and Sonoma Counties. For more information on this roadmap, please contact Evan White at evanbwhite@berkeley.edu.

[i] California Competes, *Out of the Dark: Bringing California's Education Data into the 21st Century*, May 2018, http://californiacompetes.org/assets/general-files/CACompetes_Data-System-Brief_Final.pdf, p. 1.

[ii] See, for example, *The Lifting Children and Families Out of Poverty Task Force Report: Recommended Strategies to Address Deep Child Poverty and Child Poverty in California*, Nov. 2018, https://www.cdss.ca.gov/Portals/9/CalWORKs/AB1520%20-%20Final%20Report.pdf?ver=2018-11-19-145600-677, pp. 62-64; *Rebuilding California's Juvenile Justice Data System: Recommendations to Improve Data Collection, Performance Measures and Outcomes for California Youth*, Jan. 2016, http://www.bscc.ca.gov/downloads/JJDWG%20Report%20FINAL%201-11-16.pdf; *Early Childhood Education in California*, Mar. 2019, https://gettingdowntofacts.com/sites/default/files/GDTFII_Report_Stipek_v2.pdf, pp. 215-30.

[iii] We are encouraged by the Governor's budget proposal and trailer bill language regarding Cradle-to-Career data, as well as Senate Bill 2 (Glazer) and Assembly Bill 1466 (Irwin), which would each establish an education- and workforce-focused longitudinal data systems.

[iv] Data governance and data management are sometimes conflated. In reality, data hosting can be, and often is, separated from decisions about how the data is used. This arrangement works so long as the entity charged with housing the data has no stake in the approval decisions, and carries them out faithfully. Hence our preference for an independent linkage hub that is not associated with any data contributor.

[v] 44 USC § 3581, enacted on 1/14/19 ("The head of an agency shall, to the extent practicable, make any data asset maintained by the agency available, upon request, to any statistical agency or unit for purposes of developing evidence.")

[vi] See our recently published white paper on strategies for administrative data linking: https://www.capolicylab.org/linking-administrative-data/.