Early Childhood Integrated Data System Best Practice Guidance and Recommendations

As Presented by Bardic Systems, Inc.

Table of Contents

Overview 4
Methodology 4
Ecosystem Visualization 7
Data Model Development Primer 8
Conceptual Data Model Development Process 9
Conceptual Data Model 10
Foundational Entities 10
Key Early Childhood Entities 10
Logical Model 11
Data Dictionary Overview 12
Key Issues 13
Identifiers 13
The Other Horizontals 13
Student IDs for Early Childhood 14
Other Entity Identifiers 16
Assessments 16
Organizations 16
Family 17
Linking Assessments to Outcomes 19
Rubric-based 19
Standards-based 20
Normative-based 20
Cohort-Comparison-Based 21
The Hybrid Approach (Our recommendation) 21
Example Assessments in the California Early Childhood Ecosystem 22
Federated vs. Centralized Data Storage Models 24
Rationale 24
Data Ownership 25

The Need for Longitudinal, Operational and Unitary Data 25
Making the Data Useful to the Data Providers 25
Using Unitary Data 26
Longitudinal versus Operational Data 27
The Temporality of Data Sets in a State Longitudinal Data System 28
Periodicity 28
Timeliness 29
Persistence 29
Standardized Time Sets 30
Time Metadata 30
Early Childhood Integrated Data System Recommendations 32
Policy Recommendations 32
Data System Recommendations 33
Note 34
Appendix A: Resource Allocation 35
Table 1.1 Designing the Early Childhood Integrated Data portion of the State Longitudinal Data System 36
Table 1.2 Supporting the Early Childhood Integrated Data portion of the State Longitudinal Data System 38

Early Childhood Integrated Data System Best Practice Guidance and Recommendations

Overview

This document contains a variety of types of information. It begins with a visualization of the ecosystem in which an early childhood longitudinal data system would operate, a primer on how to understand the conceptual and logical models being proposed. It then moves on to a data dictionary primer, which guides the reader in understanding and using the early childhood data dictionary we have provided. Two other documents provide background related to this guidance document. They are:

1. Conceptual_Data_Model_for_SLDS_ECIDS.pdf
2. Proposed_Logical_Data_Model_for_SLDS_ECIDS.pdf

The next section addresses some key issues in constructing a State Longitudinal Data System (SLDS) that includes early childhood data and some “on the ground” issues to deal with.

The document then concludes with a look at some possible preliminary resource allocation for both designing and sustaining the California Early Childhood Integrated Data Systems (ECIDS).

Methodology

This work was commissioned by the Santa Clara County Office of Education (SCCOE) as part of the initiation of a California Early Childhood Integrated Data System (ECIDS). Three important sources of information were used to form the recommendations contained in this document.

First, an analysis of other state approaches was used to form a baseline of information about Early Childhood Integrated Data Systems.

Second, webinars conducted by the California Department of Education on the subject of Early Childhood Integrated Data Systems were used to guide the consideration of various issues with respect to the recommendations contained in this document. Below is a list of the three webinars along with the dates the webinars were presented and a description of the presentation:

1. **CA ECIDS Webinar 1: *What is an ECIDS?: An overview of early childhood integrated data systems.* Thursday August 22, 2019.**

   Description: An early childhood integrated data system is a tool that can support state and local decision-making across programs. Participants will learn the definition of an ECIDS, understand why California would want an ECIDS, and what California should consider when exploring the creation of an ECIDS. Representatives from Minnesota and North Carolina will provide examples of how the local and state agencies work together to create a vision for the ECIDS.

2. **CA ECIDS Webinar 201: *What can you do with an ECIDS?: Lessons learned when using data to inform program and policy*. Friday September 13, 2019**

   Description: An ECIDS is designed to inform program and policy decisions that span any one program or agency. On this webinar, participants will learn about the various uses of an ECIDS from other states and what California would want to consider when considering what an ECIDS would do for the state. Representatives from Utah, Georgia, and Texas will share their experience with turning data into information.

3. **CA ECIDS Webinar 301: *What does it take to create an ECIDS?: Technical design & Privacy lessons learned*. Friday October 18, 2019**

   Description: There are three approaches to designing an ECIDS that vary based on key considerations within a state. Participants will learn about these approaches and the program and privacy implications for California to consider. Representatives from other states will share how they made their design decisions and the resulting implications for how state and local partners can use the data.

The third source of information used to form the recommendations in this document was the input and feedback from at least ten California counties. A series of three webinars was conducted in which questions were posed to these stakeholders. Comments,

questions, and feedback were collected. The webinars were conducted on the following dates:

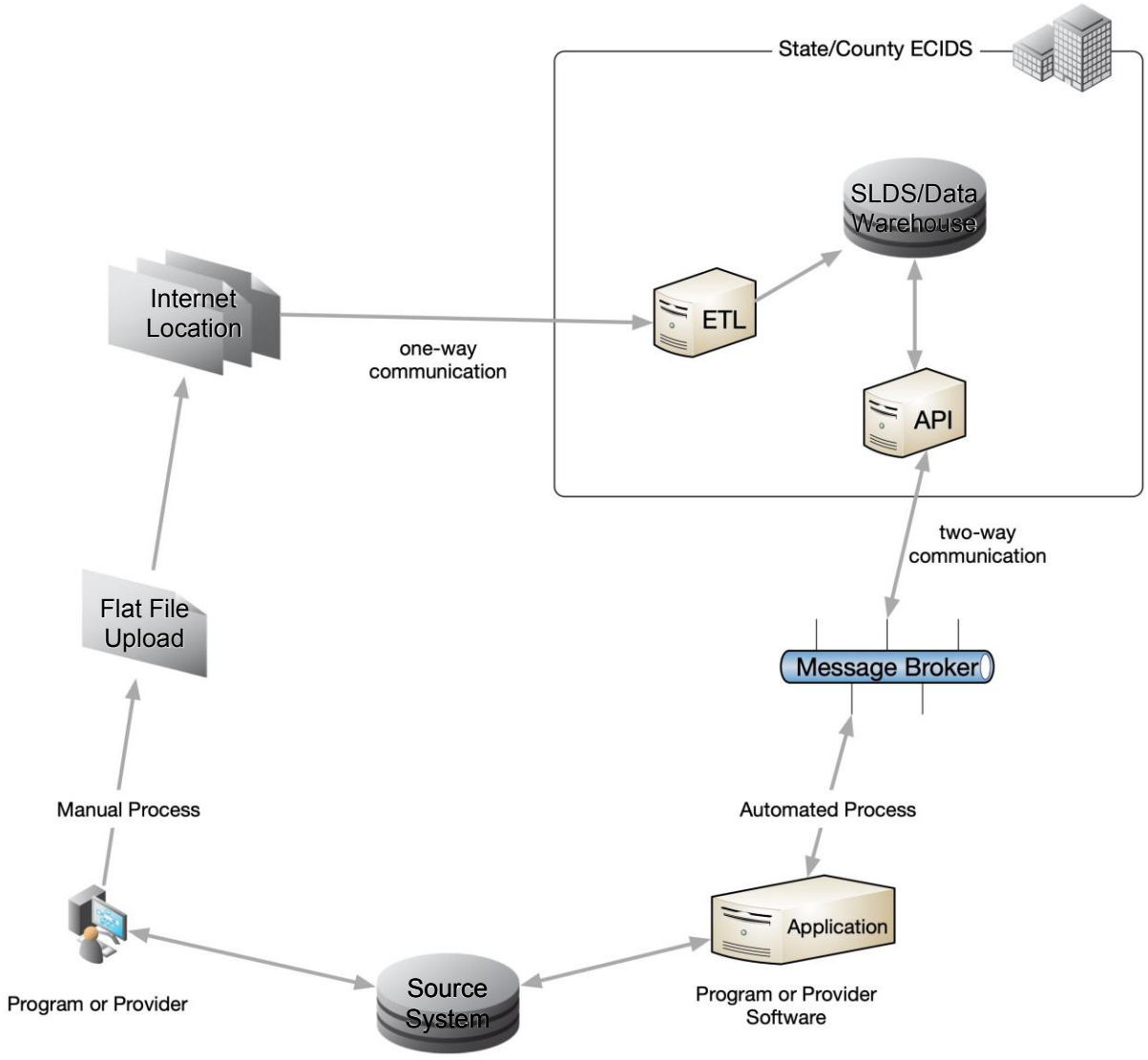
- Local Early Childhood Integrated Data System (ECIDS) Stakeholders Webinar #1:
  Monday November 4, 2019
- Local Early Childhood Integrated Data System (ECIDS) Stakeholders Webinar #2:
  Friday November 15, 2019
- Local Early Childhood Integrated Data System (ECIDS) Stakeholders Webinar #3:
  Wednesday December 4, 2019

A document[1] produced as part of a related project named *Local Early Childhood Integrated Data System* summarizes and analyzes the insights gained from the webinars.

[1] Gold, Richard (2019), *Local Early Childhood Integrated Data System.* Santa Clara County Office of Education under contract to the California Department of Education.

Ecosystem Visualization

The diagram below is a visualization of a possible ecosystem (or enterprise architecture) for a state-wide Early Childhood Integrated Data System (ECIDS). The diagram is a high-level simplified depiction of the major components and how they connect with each other. It is not meant to be a design artifact but rather to communicate a possible direction for the system. The diagram is similar in intent to the state-level Early Childhood Integrated Data System (ECIDS) diagrams presented in the NCES document *Which ECIDS System Model is Best for our State ECIDS?*[2]



[2] Duarte, S., Sellers, J., and Cochenour, M. (2014). Which ECIDS System Model is Best for our State ECIDS? U.S. Department of Education. Washington, DC: National Center for Education Statistics.

The ecosystem depicted is a hybrid of two approaches. One approach uses a centralized system in which information is collected from early childhood programs and providers and then consolidated into a database at the county or state level. The other parallel approach is a distributed system in which the centralized system collects data in a real-time, message-based manner. Although the centralized system may have a copy of some of the data in the program and providers' systems, the unitary data[3] is owned by the programs and providers, not the centralized system.

Data Model Development Primer

A comprehensive development of the data model portion of a software system is important in ensuring that the resulting system fulfills the needs and desires of the stakeholders in the project. Often overlooked, or relegated to the end of the project with little staff support, an insufficient effort in data modeling can lead to backtracking, rewrites, and a system that does not address all of the original requirements for the software system. Also, besides the user interface, it is the best way for non-technical stakeholders to communicate system requirements.

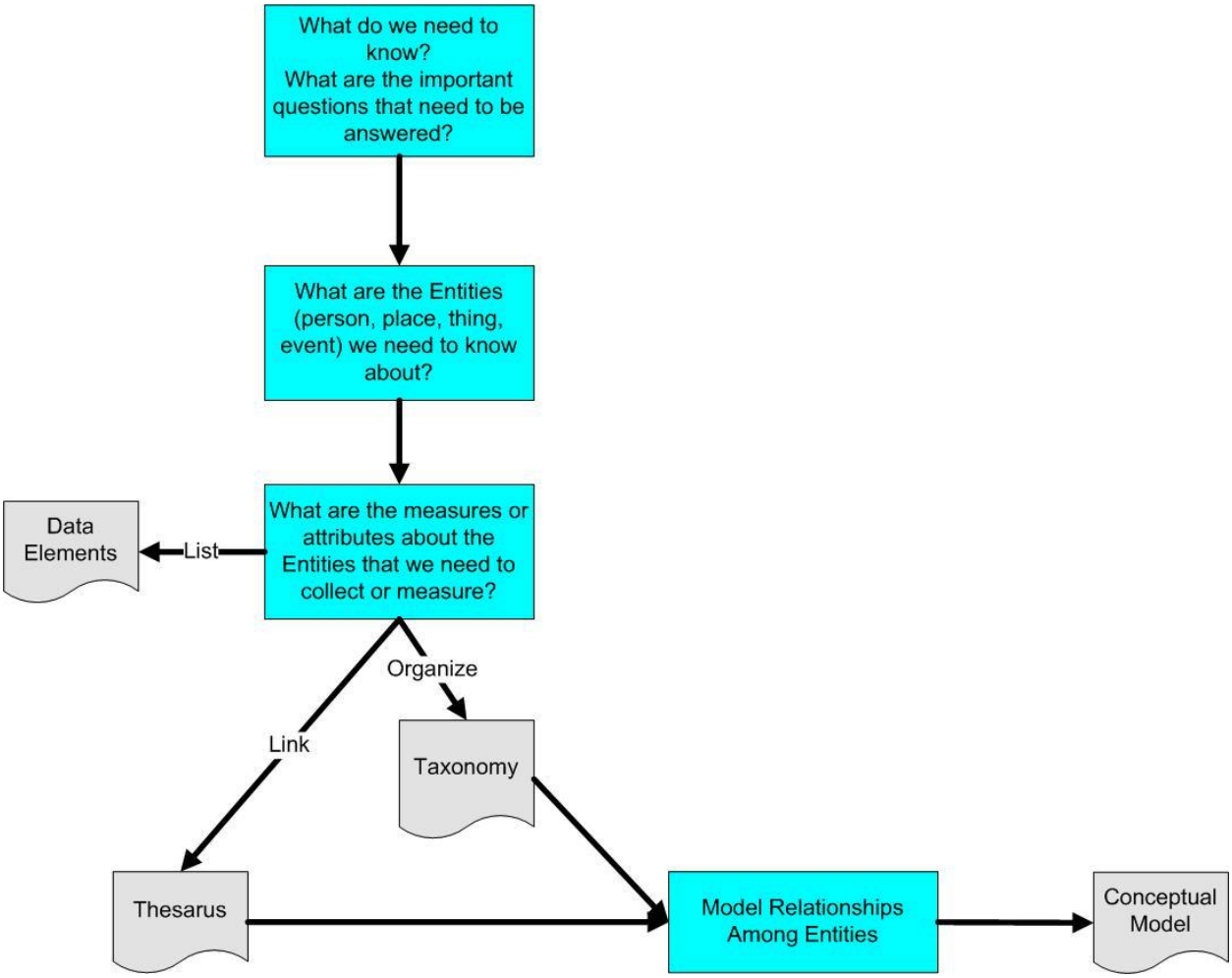
The eventual goal of data model development is the physical data model, implemented in a data management system (dbms). However, two prior steps to the development of the physical model have been accepted as part of the art and science of data model development, i.e., the development of **Conceptual** model and the development of **Logical** model[4].

[3] Unitary data means information in its most detailed and granular form. It is unaggregated data that has not been averaged or summed. It may be deidentified but a unitary piece of data (or unit record) represent a single entity such as a child, an adult, or an organization.

[4] American National Standards Institute. 1975. "ANSI/X3/SPARC Study Group on Data Base Management Systems; Interim Report". FDT(Bulletin of ACM SIGMOD) 7:2.

Conceptual Data Model Development Process

The conceptual data model can be developed using the general process depicted below. This process should be viewed as iterative in that parts of the process may be repeated in order to refine the model.



It is very important that end users, subject matter experts, and other stakeholders participate in this process.

Conceptual Data Model

Foundational Entities

These foundational entities have been developing in the Common Education Data Standards (CEDS) workgroups recently. They are very useful in guiding the development of data models in the education domain and many other domains as well.

Key Early Childhood Entities

As a result of the Santa Clara County Office of Education data model development process, and with the influence of the Common Education Data Standards (CEDS) Early Childhood data elements, the following entities were identified as key entities. We believe that this list of entities should have wide application in any integrated early childhood environment.

- **Adult** - A fully developed person from maturity onward.
- **Child** - A young person of either sex.
- **ClassGroup** - A body of students who are taught together. This entity represents a thing also known as an instruction group, and class.

- **Client** - A child or adult that receives services.
- **Facility** - A location or building in which services are rendered or administration of the program is done.
- **Family** - A social unit living together: includes Adults, Children, Guardians and other relations. Also known as household, and sibling group.
- **Group** - Any number of entities (members) considered as a unit. The group is formed for instruction, analysis and other reasons.
- **Interaction** - An in-person or virtual encounter between a staff service provider and a client for the purpose of instruction or the delivery of services.
- **Organization** - Elements related to the Organization entity. Organizations have programs that deliver services or instruction. Organizations can be early learning providers, schools, districts, and other providers.
- **ParentGuardian** - A parent/guardian is an adult responsible in some way for a child. A father or mother; one who gives birth to or nurtures and raises a child; a relative, or other adult, who plays the role of guardian.
- **Program** - A system of projects or services intended to meet an educational need. This only includes elements about the program and not child participation, eligibility, and the like.
- **School** - A type of organization in which clients are provided education services.
- **Service** - Work done by one person or group that benefits another.
- **Staff** - The body of teachers, administrators, and others that work at a school. A staff person interacts with clients.

Logical Model

The logical model moves from the identification of entities and their relationships (conceptual model) to the identification of entity attributes, i.e., data elements, and the structural relationships among the entities. This means that the logical model should be close to a physical model in terms of structure but should not yet contain Database Management System (DBMS) platform-specific information or accommodations. Platform-specific details are a characteristic of the physical data model. The logical model takes a step toward implementation but still provides for discussion and input from stakeholders.

A separate document describes the logical model.

Data Dictionary Overview

A data dictionary in the context of the Santa Clara County Office of Education Early Childhood Integrated Data System (ECIDS) project is a categorized list of potential data elements for the Early Childhood Integrated Data System. The initial list was heavily influenced by the Common Education Data Standards (CEDS) early childhood elements. The list will be modified based upon local availability and needs.

A data dictionary is a key tool not only at the beginning of a project but also through design and implementation. The data dictionary contains comprehensive metadata (data about the data). Typically, data dictionaries contain format and type information such as field length, whether the data element should be stored as character, or numeric type, etc.

This early childhood integration project data dictionary, however, also contains information on the provenance, location, and availability of each data element. Not all of this information will be available, but the data dictionary provides for tracking of this important information as integration of data takes place. The term “integration of data” belies the multi-faceted technical, legal, and community effort needed to accomplish the task. The data dictionary helps elucidate and track the status and progress along these lines.

See the *ECIDS_Elements spreadsheet*, Column Definitions tab for a list of the metadata currently tracked.

Key Issues

Identifiers

It is absolutely imperative that we have a statewide unique identifier for any person, organization, and entity in the Early Childhood Integrated Data System (ECIDS). The purpose of the system is to provide longitudinal data that allows researchers and state leadership to determine which programs are effective; and how we can improve education, training, and workforce placement. To do this implies that we have the ability to follow a student from their first experience in early childhood all the way through their retirement. That means we will need to, as much as possible, create IDs that can be used to either identify the person throughout their life or be crosswalked to the many IDs a person may have.

The Other Horizontals

Most workforce data solely use the **Social Security Number**. For a whole host of reasons that is not practical for children or persons between the ages of birth and eighteen.

At the university level (4-year colleges and graduate schools) either the **social security number** is used or a **proprietary, university-specific identifier**. At the community college level a California Community College ID (**CCID)** is established for each student statewide by the Open California Community College (OpenCCC) project.

At Primary and Secondary Education level, students are assigned a **State Student Identifier** that allows them to be uniquely identified across the entire state system which is not true of vendor-created Student IDs such as from PowerSchool or Infinite Campus, or local County or District IDs which are only valid within that county. As a result, it is unquestionably the right choice to use State Student Identifiers for children in Primary and Secondary schools.

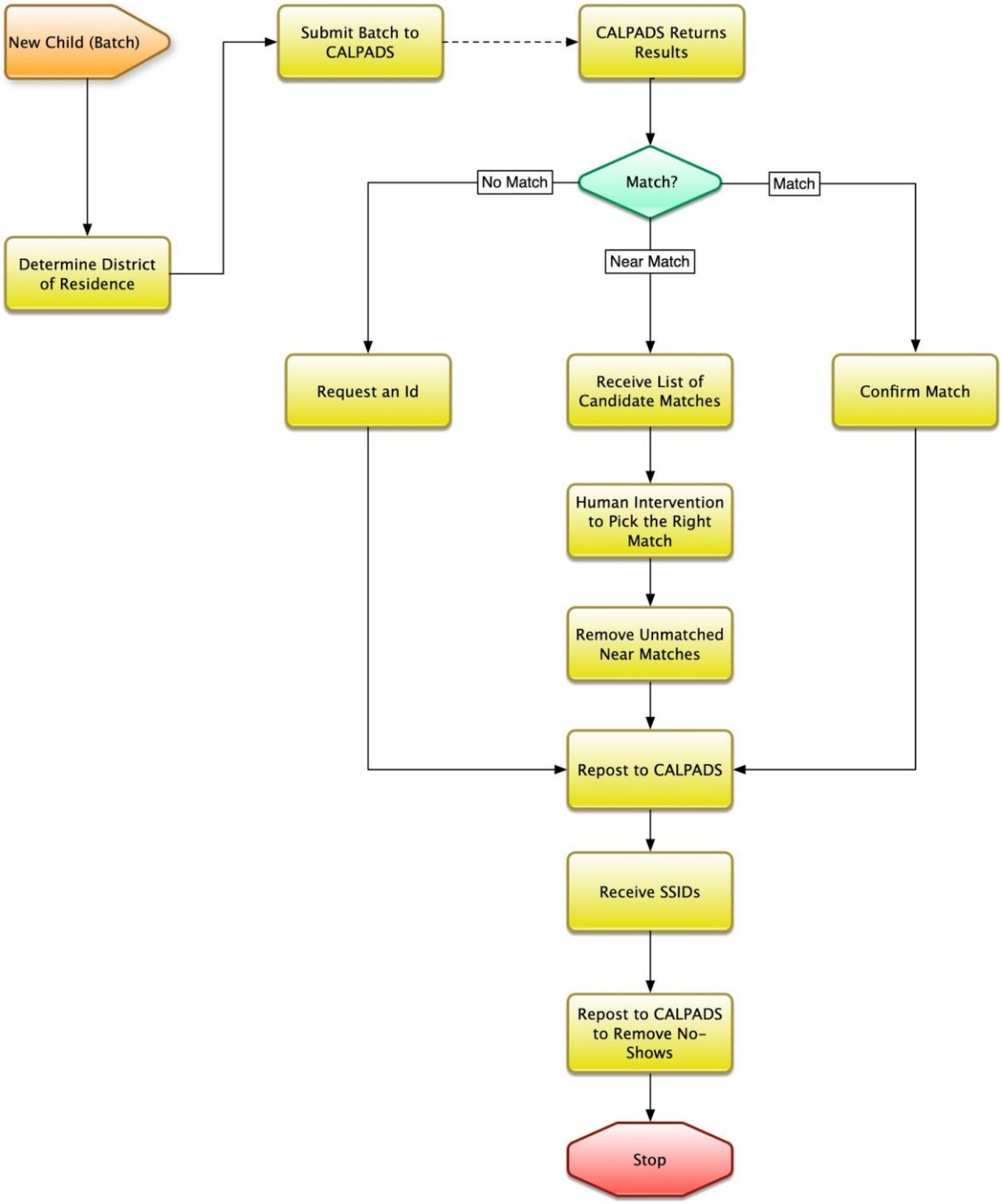
The recommendations in this document are based on the feedback we received from the local stakeholders and comparing to other states. This was strongly supported and requested during the Local ECIDS Webinars. There was overwhelming support for using the State Student Identifier (SSID) and leveraging the CALPADS system to do that.

Student IDs for Early Childhood

Since early childhood providers do NOT have a statewide identifier support structure, **we recommend that a practice be put in place that early childhood providers can and should obtain an ID from the California Longitudinal Pupil Achievement Data System (CALPADS)** — a longitudinal data system used to maintain individual-level data including student demographics, course data, discipline, assessments, staff assignments, and other data for state and federal reporting. By getting an ID from CALPADs as an early childhood program participant, every child can have a State Student Identifier (SSID), and this identifier will then follow that child through high school. The connection between the State Student Identifier (SSID) and the PostSecondary IDs is being worked out but is out of the scope of this guidance.

This is not theoretical. The **Santa Clara County Office of Education** (SCCOE) has been the central mechanism for getting State Student Identifiers and will coordinate the State Student Identifier process for publicly subsidized preschools, and currently nine major providers (San José Unified School District, Gilroy Unified School District, Santa Clara County Office of Education, Santa Clara Unified School District, Campbell Union School District, Go Kids-Estrella, San Juan Bautista Child Development Centers, Kidango and Child Development Centers/Continuing Development Inc.). Also, the **Fresno Unified School District** has been pulling State Student Identifiers weekly for children as young as 6 weeks old.

In Santa Clara County this is done by having the county office upload the required data elements from the early childhood providers and then submitting them to CALPADs which then returns a State Student Identifier for each child. The diagram below depicts a high-level view of this process.



There was a general consensus among stakeholders in the webinars that CALPADS should be the vehicle for establishing SSIDs for early learners across the state. There was also a consensus that we should broaden the capabilities of the system to allow for this.

There were some voices during our stakeholder discussions suggesting that the state and CALPADS work with counties to ease the process of acquiring State Student Identifiers for programs and providers in the county. This suggested API could use existing functions in CALPADS and possibly mimic web-based functionality.

However, San Mateo, Santa Clara and Fresno, for instance, are using non-API processes very successfully to acquire SSIDs. This process should be little work for CALPADS and will not require the storage of student data in the CALPADs system to accomplish this.

Other Entity Identifiers

There are other entities that will require identifiers in the State Longitudinal Data System (SLDS). This includes Programs, Classrooms/Sections, Ad Hoc Groups, After-School Activities and many others, but we will focus on three primary ones:

- Assessments
- Organizations
- Family

Assessments

We suggest that a system-wide method of uniquely identifying assessments be developed and the early childhood portion of the model should follow suit.

Because one of the ways we use to assess student success is assessments of many kinds we probably need to include them as a primary entity which will require us to uniquely identify them as well.

Organizations

Organizations need to be represented in a State Longitudinal Data System (SLDS) in a variety of ways. One is that the State Longitudinal Data System (SLDS) must be aware of when an organization was active and had students and teachers and other metadata

associated with it. Therefore, its opening and closing dates are very important. It must also have a persistent
unique identifier that does not get retired and persists for as long as the data that is to be analyzed is in the system. It must also be able to make the connections when the name of the organization or its address changes, or if there is a gap in which it was not active (e.g., it closes for a year and then reopens).

For early childhood, this becomes vital as many early childhood providers are small (perhaps one stay-at-home mother with four kids for which she is providing services). Sometimes they are not public institutions, yet they are receiving financial or other services from the county or state and must be tracked.

Instead of inventing a new mechanism for tracking these organizations, we recommend that any licensed or service-receiving early child provider receive a Program or Organization ID the way the K-12 system assigns them currently. This could be done through the County offices or the Districts.

Family

One of the key entities that is necessary to understand the student — particularly early childhood students — is the Household or Family.

Although it would be ideal to have statewide family ID, there are many issues with this:

1. There isn't one, so it would take a massive effort to instantiate one across the entire state. It would be a whole new collection and storage headache for many many vendors, the state itself, as well as the counties and schools.
2. If we created a statewide family registry, we would need to determine who owns the registry and who is allowed to access it.
3. We would need to determine which relationships between people constitute family:
    a. Step-brother?
    b. Great-aunt?
    c. Post-divorce second cousin?
4. We would need to be able to tag legal guardians and have the system be flexible enough that if an "incident" occurred and someone had to have their legal guardian status removed quickly it could be done reliably and very fast. This makes it a not great candidate for a state function.

5. There would be a fairly strong negative reaction among privacy and small government advocates against the data collection and upkeep that would be necessary to support this data set.

As a result of all these reasons we are recommending, although we are representing "Family" as an object in the logical data model, that instead we create the Family in the physical data structure as a **list of roles** that can be represented as singular relationships. Each relationship can have a *start date-time* and an *end date-time* (you are always someone's mother but NOT always their legal guardian). For instance:

1. Sally (with State Student Identifier) has a relationship <is daughter of> with Mary (with SSN or local ID)
2. Sally has relationship <is daughter of> with Frank
3. Mary also has relationship <is legal guardian of> with Sally
4. Sally DOES NOT have relationship <is legal guardian of> with Frank
5. Frank does have relationship <is secondary contact of> with Sally
6. Sally has relationship <is sibling of> with Bobbie, her older brother

This will take multiple entries to describe a Family. To find a whole family, you would need to pull all relationships with Sally and construct the Family that way. But this way you can have discrete control over what gets collected. For example, if "Frank" wants or needs to have all his information removed from the system it doesn't mean you need to break up the whole family object you just remove the relationships with him in it.

It also has the advantage that the state doesn't have to propose a singular culturally-specific definition of what a family is. It provides flexibility without losing functionality. Each relationship would have an internal ID but the relevant IDs would be the ones associated with the people. An advantage of this recommendation is that in existing local data systems, many schools and programs already have singular family relationships identified and stored.

If this is not acceptable, our backup recommendation would be to use locally-managed Family groups at the county, school, or program level. This would require near-real time updates to the centralized system. The infrastructure to support this would be extraordinary, but **VERY** expensive, and challenging to maintain.

Linking Assessments to Outcomes

This is a domain far larger than early childhood. “Assessments” does not merely refer to high stakes summative assessments but also diagnostic assessments, benchmark assessments, certifications, and components of formative instruction. Much of early

childhood assessment is developmental evaluation or components of formative instruction. Because they are so locally modified, we will not focus on those although they will need to be captured in the longitudinal system.

Another issue in early childhood is that it is difficult to rely on written assessment instruments because many of the children do not have a sufficiently developed skill in writing to be able to express their Knowledge, Skills, Aptitudes and Capacities through that medium.

The key will be to build a taxonomy to make assessment comparisons be apples to apples comparisons — which can be very challenging even with the same assessment year to year.

There are four primary methods of mapping assessments to outcomes. This allows the generation of what some call “authentic assessments”. The four methods are:

1. Rubric-based
2. Standards-based or Criterion-based
3. Normative based with a limited set of assessment instruments
4. Cohort-comparison based (Student Growth Percentile)

Rubric-based

This method involves pushing all results into a “comparative” taxonomy. One of the most common is below:

- fails to meet
- partially meets
- meets
- exceeds

Letter grades are another common example of this type of assessment. This is the most common way of generating assessments but is also the least comparable because the “grading” of these assessments, unless very carefully calibrated, can be extremely subjective.

The key is to develop a common structure at the assessment-level of outcome for a grade level, age or subject area scope. In this model you measure outcomes by selecting cut points in the assessment results that represent the outcome goal.

At the level of three or four categories the comparability of the assessment “score” is much more comparable than at a finer grain.

Standards-based

Standards/competency/criterion-based methods are the preferred method. That is because building out Knowledge, Skills, Aptitudes, Development, and Socio-Emotive Capacity as benchmarked standards allow learners to be assessed in an apples-to-apples way.

The challenge here is to find tests that can be executed in a way that does not rely on the student’s writing skills (which may be non-existent) and can be executed in a fair, and repeatable way without disrupting the teacher and student’s learning process.

In an ideal world ,a collection of output data from formative instructional components that are aligned to standards would build the student’s “learning record.”

Normative-Based

This type of assessment involves setting up a particular kind of test per grade/subject_area/ school readiness rubric and adopt that across the system. It could be a vendor — or publisher-provided assessment or a state function. This is challenging in such a distributed and federated system but some assessments — like DIBELS or QRIS — are relatively commonplace.

The challenge with this method is being able to do comparisons to other assessments structures and year-to-year variance in the assessments. The advantage is that a vendor-supported assessment can be universally applied much easier if political and policy agreement can be reached between providers. The other challenge is that his kind of scoring has limited longitudinal comparability and thus belies the longitudinal purpose of the State Longitudinal Data System (SLDS).

Cohort-Comparison-Based

For longitudinal studies student growth percentile methods produce truer representations of authentic assessment data than single point in time scores.

By using the longitudinal nature of the datasets it will be possible to create very powerful cohorts of like students and then by using only somewhat complex statistical methods a “student growth percentile” (SGP) can be calculated which defines a deviation from a norm set by a like group of students. What this does is it allows you to filter out biases like socio-economic status, race, sex, and geography by creating a “like cohort”. We would suggest using the Colorado SGP Method of deriving the Student Growth Percentile. One of the other powers of this method is it also filters out variance in the assessment instruments year to year and can be used to link the cohort and the outcomes of that cohort against any assessment instrument or outcome measurement.

The negative aspect of using SGP is that it requires two complete data sets on each child to start producing authentic value. As a result it is extremely effective when used longitudinally but does not provide much feedback for the teacher until that three terms or three years of data are in place. This makes it VERY challenging in the early childhood domain.

The Hybrid Approach (Our recommendation)

The key is to know what outcomes you want to measure against and to remember assessment instruments are only guides and you need multi-factor requirements. You cannot just look at a single score and make judgements.

At this age scoping by development milestones is important, and you may not have enough data longitudinally to provide cohort-comparisons. As a result, representing your outcomes as standards in Knowledge, Skills, Aptitudes, Development, and Socio-Emotive Capacity that the student can master allows for assessments that provide immediate feedback **AND** can be made longitudinal for future years of support for that learner. While using a particular assessment instrument or instruments across the whole system would provide some excellent comparability that may or may not be possible given the legislative and local control environment.

Example Assessments in the California Early Childhood Ecosystem

There are though common assessment instruments used across much of the California early childhood ecosystem. These are excellent candidates for standardization and statewide implementation given their already heavy use:

- **IEP Evaluations**
- **ASQ** - *Ages and Stages Questionnaire*-developmental assessment of a child; young child assessment at different ages and different milestones; screening tool from a caregiver
- **ASQ-SE** - *Ages and Stages Questionnaire-Social and Emotional* (health screening)
- **EPSDT** - *Early and Periodic Screening, Diagnosis and Treatment*-Physician
- **Early Intervention Screening** - Speech, physical, occupational
- **DRDP** - *Desired Results Development Profile* (primarily completed by a teacher or school staff person working with the child, after consulting with parents and other staff, over time and observation of skills across all developmental areas of the child; a continuum of skills and aptitudes. A quantified rating of the skills; child level assessment: one for an infant/toddler (IT); one for pre-school (PS); one for kindergarten (K). Not a health screen.
- **QRIS** - *Quality Rating and Improvement System*. State legislation for California counties to provide QRIS ratings of early child care/learning sites. Seven (7) mandatory elements: whether completed or not, to which quality in each category. Organization or classroom-level assessment: staff qualifications; quality of staff-child interaction; health and safety. 1-5 score; incentive if you score above a 3.
- **Head Start** - Has assessment data of various types.
- **Resource and Referral -** (R&R)
- **Self-Assessment** - Any data elements not covered by any of the others.

As you can see, there is a surprisingly large body of assessment instruments in use in Early Childhood. By normalizing how these assessments are stored, there is a wide variety of reporting and data mining options that could be used to generate effectiveness studies and new approaches.

Once the overarching design of the entire State Longitudinal Data System (SLDS) starts to take shape at the state level it will be possible to design the suite of assessment normalizations and taxonomic structures.

Federated vs. Centralized Data Storage Models

Currently, we recommend that the system utilize a Hybrid system with the following characteristics:

- Some source systems will look like they are Federated because they will be loosely bound to the Early Childhood Integrated Data System (ECIDS). That is, their data will not reside in the centralized data warehouse but will be available in real-time (or near real-time) to the Early Childhood Integrated Data System (ECIDS). Snapshots of the Source systems may be taken in order to provide for official statistics (numbers that don't change every day).
- Other source systems will look like they are Centralized because they will be tightly bound to the Early Childhood Integrated Data System (ECIDS). That is, their data is linked and a copy of the data is stored in the Early Childhood Integrated Data System (ECIDS); similar to what the DataZone application in use by Santa Clara does now.
- A system of identifiers will be adopted among all the source systems that will ease the ETL process for both centralized and distributed source systems. If the same identifier exists in all or most of the source systems, matching will be much easier and accurate. Please see the Identifier section of this Paper.

Rationale

The state of California is well positioned to implement a hybrid system because they have already done much work in creating institutional trust[5] among the various early childhood agencies. This will accommodate a system that has centralized as well as distributed components. Source systems that have a high level of technical expertise can keep control of their data, and still share it with the central system through a loosely coupled (distributed) component. Source systems with less technical expertise or resources might choose to provide a copy of their data to the central system thereby putting more relative responsibility for the data on the central system.

In a local example, the DataZone project in Santa Clara County Office of Education has a good centralized system on the ground and working. The loosely coupled components could be added to the existing system making it a hybrid system. This model is a powerful one for us to guide the work of the State Longitudinal Data System (SLDS) statewide.

The Santa Clara County Office of Education has used the guidance provided by the California Department of Education (CDE) Early Childhood Integrated System webinars,

[5] Presentation given at the Alameda Juvenile Justice Forum February 6, 2019, Marcy Lauck, Director SVRDT & SCCOE Data Governance. http://www.acgov.org/probation/documents/DataTechnology_MarcyLauck2.3.19.pdf

discussed previously, as well as the National Center for Education Statistics' tool kit[6], and, more specifically, the guidance concerning system model types[7].

Data Ownership

Early Childhood programs want to be sure that the data they share with the Early Childhood Integrated Data System (ECIDS) will be used appropriately and by the proper persons. Shared data may remain physically in the program offices or in a centralized data warehouse.

Alternatively, shared data may reside in a virtual environment such as the cloud under the control of either the Early Childhood Integrated Data System (ECIDS) or the program office.

This kind of intra-agency data sharing requires institutional trust and formal agreements among the agencies.

Two separate but related issues need to be considered when choosing among federated, centralized or hybrid system models. First, is the physical location of the source system data. Source system data may be in servers local to the agency, in a cloud, or exist as files in digital or paper format.

Second, is the ownership of the data and copies of the data. The Santa Clara County Office of Education recognizes that while the physical location of data is important, more important are sharing issues such as data ownership, data stewardship, control responsibilities, liability, security, intra-agency trust, and communication in determining the system architecture. This is especially important when sensitive or personal information is shared with the Early Childhood Integrated Data System (ECIDS).

The Need for Longitudinal, Operational and Unitary Data

Making the Data Useful to the Data Providers

That is, the higher authority can mandate data collection, provide resources, provide technology solutions, etc. but if the local submitter does not get the data back or does not get their own data back in a timely manner, the quality of the data will suffer. This was unanimously called for in the stakeholder meetings across the state as we discussed this topic. Local stakeholders want to see the data returned to them so they

[6] Coffey, M., Chatis, C., Sellers, J., and Taylor, R. (2014). SLDS Early Childhood Integrated Data System Guide. U.S. Department of Education. Washington, DC: National Center for Education Statistics

[7] Duarte, S., Sellers, J., and Cochenour, M. (2014). Which ECIDS System Model is Best for our State ECIDS? U.S. Department of Education. Washington, DC: National Center for Education Statistics.

can make operational use of it. This adds some complexity to the project to enable this but will VASTLY increase adoption, data quality, and continuous use.

The state comparison webinars referenced in the Methodology section discussed North Carolina's use of feedback to the local stakeholders and the value of that for usage and data quality and we should take that as a guideline. On top of that if the data is to be useful it should be diagnostic or formative data in the short term, or it will not be operationally useful to the educator in real-time or near real-time and will be much more accurate, and clean.

Using Unitary Data

For data to be able to be analyzed across horizontals from early childhood all the way to workforce data it must in the end be Unitary data and contain identifiers so that each record can be linked to the person they are associated with.

This is important and often counter to common thought because it does raise privacy and security issues. However, there are several reasons why it is critical:

1. As long as the data is protected by obfuscation policies so that the Personally Identifiable Information (PII) or identifiable data record groups are not publicly released that covers much of the privacy concerns.
2. For the data to be able to be vetted by the local source for quality control AND for it to be useful to the local source it must be unitary data and contain identifiers that are meaningful to the local source.
3. Family relationships and geo-data about a learner changes relatively rapidly and not to be able to follow a student or Family as they move around the system would impair some of the utility of the State Longitudinal Data System (SLDS).
4. One of the uses of a system like this is to give educators the ability to compare immediate assessment results with data on the student even if the student was in another county or school. The moment you stipulate that the system should be tactically useful to educators it requires quick turn-around and unitary data.
5. We don't know what we will need to know once we have two or three years of data from the system — we may discover trends and insights that then force us to ask questions we didn't expect and if all we have is aggregated data we will need to redo all the data collection from every source in the state which would be painful, expensive, and a politically costly process.

Although it critical to have unitary data, privacy and security concerns are paramount to deal with. There are many questions that would need to be addressed both in outreach and in policy:

1. Who gets to see unitary records?
    a. Current teachers? Former teachers? Future Teachers?
    b. Can legal guardians see data from before they were the legal guardian?
    c. Can they see data that existed while they were a legal guardian after they aren't one anymore?
2. Who can manage the data store with the unitary records?
3. What is the state's obfuscation policy (integer rounding, top-down and bottom-up rounding, n-levels)?
4. What data does the state consider to be PII and thus must have extra protections?

Longitudinal versus Operational Data

There are some design considerations if we want to make the State Longitudinal Data System (SLDS) have any operational capabilities. This is important to consider given the strong feedback from the field that this is critical for their participation.

One is access activity. Most longitudinal data structures are optimized to Read data very quickly, and not optimized for data entry. This also called "enter once, access many."

Operational data structures tend to be optimized for transactional activities based on the limited use cases they are designed for. If it has to do BOTH then that means you need to think very carefully on design — either the centralized data architecture itself or the decentralized service infrastructure to make sure they can serve those purposes. This will have an impact on both time to implementation and cost.

The other issue with longitudinal data is that to have broad, high-grain comparisons you need fixed time sets with a fixed data scope. The time sets and temporal metadata is discussed in the Temporality section of this paper, but the data scope is important as well.

Data in the system must be comparable year to year. Invariant data structures are very important both in structure and the attached metadata. Any sub-grouping or filtering you

can imagine doing on the statewide data sets has to be accounted for in the data record. As an example, if you plan on comparing by standard assessment outcomes then you need those standard outcomes to be attached to those records every year. If you plan on doing studies on student mobility or serving transient students, you need to make sure that zip code or address metadata is always attached to demographic records. This may seem obvious but is very often overlooked in State Longitudinal Data System (SLDS) design and has profound impact on data collection burden on the counties and districts you are getting data from. You want to get this as close to right as possible the first time as it is expensive and politically charged to change the data collection demand frequently.

The Temporality of Data Sets in a State Longitudinal Data System

One of the most common organizing structures in a State Longitudinal Data System (SLDS) is TIME. This has many faces. One must be concerned about periodicity, timeliness, persistence, standardized time sets and time metadata

Periodicity

The time correlation to the data depends on how often and how frequently data is uploaded. That is called “periodicity” or the “period of collection.” It is critical that the data collection periods map to the business need to have data from particular time periods. This is another often overlooked criticality. One of the questions to ask is how often is the data collected, uploaded, or updated?

One thing to deal with is the fantasy of real-time data collection and the reality of near-real-time. It is possible to get data in near-real-time. However, there are many dependencies:

- It is dependent on how frequently the human process updates the source system. In Wyoming there was a SIF-based real-time data collection of attendance data and it would be empty all week and then after the weekend it would have the last week’s data.

  The technical staff looked and looked, and it appeared the computers were talking correctly. It turned out that the school secretary kept all the attendances — which were collected from all the rooms on paper — in her inbox until Friday

afternoon and then put them in her school's system. The human process always trumps the computer process!

- How often do the computers update each other? In many automated API-driven systems there is a gate that only opens once every five minutes or once an hour or once a day at midnight, or every Sunday at 3:00 pm. No matter how current the local system is, the State Longitudinal Data System (SLDS) will only be updated at the periodicity the gate opens up!
- In any automated system things "hang". If some component stops working, the technical support and troubleshooting needs to be in place to restore the flow of data. The more that kind of trouble-shooting can be built into the design, the more uptime the system will experience.

Timeliness

It is critical — as has been discussed elsewhere in this report — that data be made available to the users of the State Longitudinal Data System (SLDS) WHEN they need it. If you are attempting to help guide a school's decision making for this year's cohort and then deliver it in August after the students have left school, then the system has failed the timeliness test. If you are supporting teaching and learning then early childhood providers (or any educator) will need data and information about their students in time to intervene, principals or program directors in time to intervene with their schools and teachers, superintendents time to intervene with their district results, etc.

Timeliness scope should be explicitly built into the initial scope of the system dependent on the role and context of the users it is initially directed toward.

Persistence

How long data is kept is a function of the laws of the local, state and federal government. Data that is de-identified can be kept indefinitely for future research as long as the PII data is removed or obfuscated. Persistence is a key function of a State Longitudinal Data System (SLDS). You cannot generate longitudinal reports and student growth percentile with data over time.

Persistence also have a year-to-year correlation component. Some data is ONLY applicable to certain age or grade levels or subject areas. Certain data stored by early childhood providers on development level or socio-emotive competency or lexile score or quantile scores may become dated and can only be used to do program evaluation based on future performance kinds of longitudinal studies but may not be meaningful as

the learner has moved on. However, this data still needs to be persisted for long-range longitudinal studies which can have a powerful impact on future students.

Standardized Time Sets

In a longitudinal system it is critical that data be tagged with standardized time sets that will be typically used by its end-users in queries. The next section will talk about time metadata like end time and start times, but this section is about the business structures that people use to think about education. For instance:

- Academic year
- Calendar year
- Fiscal year
- Terms (Semesters, Trimesters, Quarters)
- Months
- If appropriate even section/periods.

These standardized time sets are critical to the usability of the system and must be available whether they are hard coded in (every record has an academic year associated with it) or derived from granular time metadata (start time of September 3, 2020 and end time of May 13, 2021 means it is Academic Year 20-21).

If the academic year is an important feature, then it is critical that one explicitly tag things with that as it will save having to make awkward algorithmic calculations to determine the academic year — particularly when counties might have subtly different ways of determining their start and end dates and summer sessions, etc.

Time Metadata

There are granular tags that need to be associated with records to place them in detailed time.

The common tags/columns/properties include:

- Start DateTime
- End DateTime
- Effective Date
- Inactive Date

- Duration

These tags are designed to be able to place the record in time. It is essential to always have a time context for every record/row/object so that it can be placed in time and inside a standardized time structure where they exist. It allows for questions to be answered like:

- How many of XXX happened in May of 2012?
- What is the average # of discipline incidents per month between January 2013 and July of 2019?
- How many students were enrolled in Pocarro Early Childhood Program throughout 2018? (Notice that is different than if someone used a standardized time structure they could ask “How many students were enrolled in Pocarro Early Childhood Program in Academic Year 17-18?”)

Early Childhood Integrated Data System Recommendations

Our core recommendations mirror the recommendations generated by the Early Childhood Data Collaborative (ecedata.org) after their two 2013 and 2018 surveys across the United States early childhood ecosystem. Some of these are stretch goals but they make excellent targets for the system. Many of these were included, adjusted and prioritized based on the feedback we received from the stakeholders. That feedback includes:

- The system should use CALPADs to generate SSIDs for early learners.
- An almost universal push for unitary data
- A strong desire to have the data loop back to the stakeholders in a timely manner not ANNUALLY
- Several request that they be able to input data once and have that populate the whole system rather than having to input data multiple times.

Policy Recommendations

1. Establish and support an Early Childhood Integrated Data System (ECIDS) data governance body (or bodies) to guide the coordination, security, and appropriate use of early childhood data.
2. Strengthen California's capacity to securely link data on young children across all state and federal early childhood programs, including Head Start and home visiting.
3. Expand efforts to collect and link data about the early childhood workforce.
4. Communicate with parents and legislators and policy-makers about data privacy policies and uses of early childhood data.
5. Use already existing data systems planning tools and technical assistance to support early childhood data system integration as well as any new State Longitudinal Data System (SLDS) tools. California should leverage:
    a. the Federal State Longitudinal Data System Support Team (SLDS Team- https://nces.ed.gov/programs/slds/techassistance.asp),

b. the Center for IDEA Early Childhood Data Systems Center (DaSY Center- https://dasycenter.org/),
c. the Privacy Technical Assistance Center (PTAC- https://studentprivacy.ed.gov/),
d. and join the Student Data Privacy Consortium (SDPC- https://privacy.a4l.org/)

Data System Recommendations

A fully functional statewide data system for early childhood would include:

1. Unique statewide student/child identifier (State Student Identifier).
2. A cross walk to adult and postsecondary IDs (California California Community College ID (CCCID) and, possibly, carefully, Social Security Number)
3. A methodology for linking family/case data with child data (See Family ID section).
4. Child-level demographics and program participation information.
5. Child level data on development and/or the child's school readiness (see Assessment and Outcomes section).
6. Ability to link child-level data with K-12 and other key data systems (State Student Identifier and Academic Standards Alignment).
7. Ability to link child level education data with health and family data (a great example of this is Rhode Island's KIDSNET [http://health.ri.gov/publications/factsheets/Kidsnet.pdf]).
8. Ability to link with Head Start programs (See Santa Clara Head Start coordination).
9. Unique program site identifier with the ability to link with children and the Early Childhood workforce (this will need to include private providers to be comprehensive).
10. Program site structural and quality information (see note above on including private providers and including licensing status, structural standards, working conditions, and quality measures).
11. Unique Early Childhood workforce identifier with ability to link with program sites and children (includes work records and organizational registries).

12. Individual-level data on Early Childhood workforce demographics, education, and professional development information (may require partnership with workforce or some kind of Data Trust relationship between agencies).
13. State governance body to manage data collection and use (A California Data Governance Council will be essential to manage both the privacy and security questions referenced below and questions of the appropriate metrics in both scope and grain).
14. Transparent privacy protection and security policies and practices (this will need to be far deeper than just “We will abide by law XXX”. There will need to be a systemic approach to handling these issues or the project could fail sheerly on this factor).

Note

The project described was supported by the Preschool Development Grant Birth Through Five Initiative (PDG), Grant Number 90TP0015-01-01 from the Office of Child Care, Administration for Children and Families, U.S. Department of Health and Human Services. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Office of Child Care, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

Appendix A: Resource Allocation

Until the State Longitudinal Data System (SLDS) team (referring to the overall team implementing the Governor's vision for a cross-horizontal State Longitudinal Data System) decides on their approach it is impossible to determine local resource usage. If the state funds a centralized team that would manage the data collection and data quality and verification processes then the load on Counties and Districts will be very light. This would be ideal in terms of equity of access. A strong centralized team would be able to serve both giant counties and tiny ones equally well.

Some risks and mitigations are discussed in the recommendation section above, but it is far too early to be able to even suggest vendor options, timelines and software — and hardware-based budgets of the Early Childhood State Longitudinal Data System because the scope and design of the statewide State Longitudinal Data System would need to be clarified more first. We will need more information on what the entire system will look like before we can put any detailed recommendations in place about cost and team composition.

Despite the uncertainty of the State Longitudinal Data System (SLDS) approach, we can still estimate the resource needs for designing the Early Childhood portion of the State Longitudinal Data System based on what other states have done in the State Longitudinal Data System domain. We can also make an educated guess at what a statewide support team for the entire State Longitudinal Data System might look like.

In the table below when we refer to horizontals we are referring to Early Childhood/PreKindergarten, Primary (K-6), Secondary (Middle, High School), Postsecondary (Community Colleges, 4 year Colleges and Graduate School), Military, Workforce, and Training.

Table 1.1 Designing the Early Childhood Integrated Data portion of the State Longitudinal Data System

Role	Year 1 FTE	Year 2 FTE	Description
Data Architect	0.5	0.2	This role is designed to oversee the design of the Early Childhood portion of the model and ensure that it interfaces cleanly and with no overlap with the other horizontals.
Technical Lead	1.0	0.75	The technical lead will ensure that the various technical teams are harmonized and that no team works toward the wrong objectives .
Database Developer	0.75	0.5	The Database Developer will work hand in glove with the Data Architect and Technical Lead to build out the model and ETL (Extract Transform Load) hooks appropriately inside the bigger model.
ETL Developer	0.75	0.75	The ETL developer will work with all the data sources to ensure that there are appropriate APIs and ETL structures inside the model and synchronized with shared stakeholders between horizontals.
Rules Developer	0.5	0.75	The rules developer will use whatever rules/data quality control technology decided on to ensure that the data from the sources can be quality checked and synchronized in order to be equivalent with data across the horizontals AND still operationally useful for local stakeholders.
UX Developer	0.5	1.0	The UX developer will help design the user experience of querying, pulling data from the

			system, report design, and the admin management screens of the system.
Quality Assurance (QA) Lead	0.5	1.0	The Quality Assurance (QA) Lead will have a challenging time with this project. They will need to design test cases in cooperation with QA leads from all the other horizontals and the Data Architect and the Tech Lead. The Quality Assurance Lead will also need to lead the Quality Assurance team to make sure a quality product is released under the time constraints.
Quality Assurance Team (2)	0.5	2.0	The Quality Assurance Team will need to execute against the test cases and document the execution.
Project Manager	0.25	0.5	The Project Manager will be one of a team of Project Managers from the other horizontals. This person will need to make sure that everything is synchronized amongst the various stakeholders and horizontals.

Table 1.2 Supporting the Early Childhood Integrated Data portion of the State Longitudinal Data System

Assumptions:

- This estimate assumes that there will be a centralized team supporting the State Longitudinal Data System (SLDS) project and these roles will be an adjunct to the centralized team.
- Level 1 Help Desk for the State Longitudinal Data System is also assumed to be centralized at the state level for the entire State Longitudinal Data System across all horizontals.

Role	Headcount	FTE	Description
Level II Early Childhood Integrated Data System (ECIDS) Support	1.0	0.5	This role is a part time role with half the person's time being charged against the State Longitudinal Data System (SLDS) project. This person will triage any Early Childhood support requests that come in from the Level 1 Support team. Part of their job will be to distinguish between data, technical, and policy issues.
Technical Lead	1.0	0.5	The technical lead will not be a full time resource but be a generalist that would respond to and work with the Level II support specialist with technical ETL (Extract Transform Load), data storage, rules and User Experience (UX) issues that are Early Childhood specific and not handled by the centralized State Longitudinal Data System (SLDS) technical team.
Quality Assurance Lead	1.0	0.5	The Quality Assurance Lead will be a part time resource and a generalist that will work in partnership with the Technical Lead on any changes, additions, or deletions they make. They will be responsible as well to check data lineage to make sure no changes break any of the Early Childhood State Longitudinal Data System (SLDS) functionality.
Project	1.0	0.5	The Project Manager will manage the Help

Manager			Desk Specialist, Technical Lead, and Quality Assurance Lead in their work. They will also manage reporting on the system and escalating any political, policy, or strategic issues to State Longitudinal Data System (SLDS) leadership. They will also manage relationships with Early Childhood stakeholders and act as a communication liaison to that community.