CALIFORNIA Data System

# Understanding Data Transfer Processes

March 2020 Technology & Security Subcommittee
Cameron Sublett, WestEd

In 2019, California enacted the Cradle-to-Career Data System Act (Act), which called for the establishment of a state longitudinal data system (SLDS) to link existing education, social services, and workforce information.[1] The Act also articulated the scope of an 18-month planning process to be shaped by a workgroup that consists of the partner entities named in the Act.[2] Suggestions from this workgroup will inform a report to the legislature and the designs for the state data system to be approved by the Governor's Office. The Technology & Security Subcommittee will support the workgroup by determining technology specification requirements related to data structure and security.

This brief elevates three topics for consideration in advance of the March convening of the Technology & Security Subcommittee: data ingestion, data integration, and data security. After providing a high-level description of each process, the brief highlights the work of two other states for the purpose of drawing contrasts. The document concludes with questions that the subcommittee will consider at their second meeting, in order to refine recommendations related to the data structure.

1 Read the California Cradle-to-Career Data System Act at: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=EDC&division=1.&title=1.&part=7.&chapter=8.5.&article=

2 The partner entities include the Association of Independent California Colleges and Universities, Bureau for Private Postsecondary Education, California Community Colleges, California Department of Education, California Department of Social Services, California Department of Technology, California Health and Human Services Agency, California School Information Services, California State University, California Student Aid Commission, Commission on Teacher Credentialing, Employment Development Department, Labor and Workforce Development Agency, State Board of Education, and University of California.

## Three Facets of the Data Transfer Process

In the February 2020 meeting, the Technology & Security Subcommittee discussed the trade-offs between centralized and federated data structures, how such structures might influence data alignment, recency, and security, and whether one data structure could support both a P20W data system and a data request process. The subcommittee recommended that California adopt a hybrid approach, where some information is held in a centralized location and other information is accessed only when needed. The Cradle-to-Career Workgroup supported the concept of a hybrid approach and requested that the subcommittee determine which elements should be in the P20W data set and which items should be accessed using a federated model. In addition, the workgroup requested that the subcommittee clarify how cloud storage technologies would be utilized in both centralized and federated contexts.

The subcommittee will address these topics in their March gathering by taking a sharper focus on three technical topics:

1. Data ingestion
2. Data integration
3. Data security

Data ingestion, integration, and security are parts of a larger, more complicated data transfer process that will need to be specified prior to creating the state data system. Stated in plain language, a data transfer process details how information comes into and out of the data system. This includes the technical processes of exporting, importing, processing, storing, maintaining, and making data available. Figures 1–3 illustrate the data transfer processes used in Oregon, Washington, and Utah.

Figure 1: Oregon's data transfer process[3]

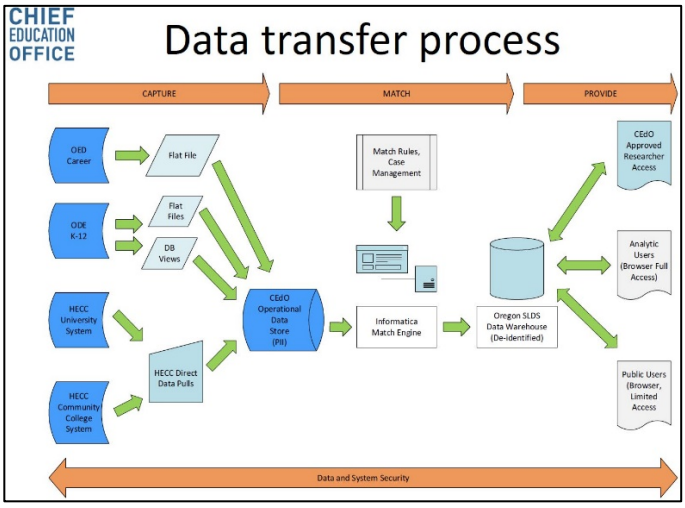
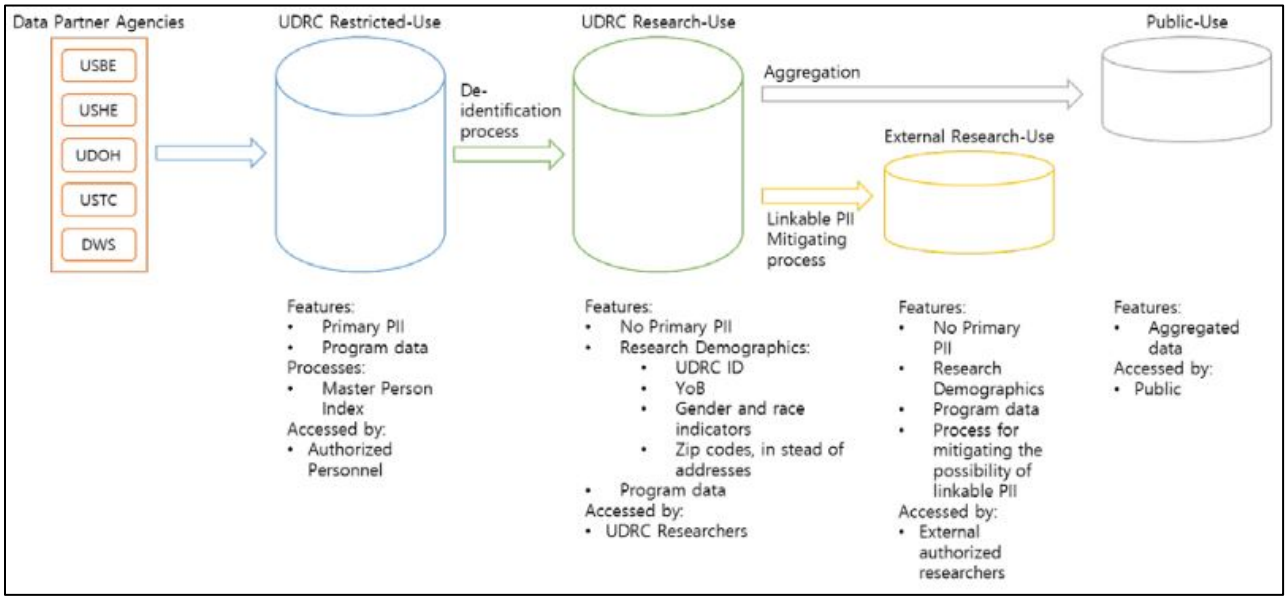



Figure 2: Washington's data transfer process[4]


3 See the Oregon data transfer process here: https://www.oregon.gov/highered/research/Documents/SLDS/SLDS-Data-Flow.pdf

4 See the Washington data schematic here: http://slds.rhaskell.org/state-profiles/washington

Figure 3: Utah's data transfer process[5]




## Data Ingestion

The data transfer process begins with data ingestion, which refers to importing, transferring, loading, and, to some small degree, processing data for storage or analysis. In the context of an SLDS, data ingestion refers to the intake of raw data from partnering education or workforce entities. Processes and protocols related to data ingestion vary.

Utah has a centralized SLDS,[6] and according to Jeremias Solari, Assistant Director of the Utah Data Research Center (UDRC), data ingestion in Utah is facilitated by an in-house, exportable, Java-based application that partnering entities use to funnel raw data into storage. It works by specifying a Microsoft SQL statement, encrypting data, and exporting the data using a Secure File Transfer Protocol (SFTP) server. Data are then imported, algorithmically matched, stripped of all personally identifiable information (PII), and stored in a researcher database that is only accessible by UDRC research personnel. UDRC data are batch-processed on an annual basis. These data are physically stored in a building near the Utah state capital, though UDRC is currently weighing the use of Google Cloud Storage in the future. UDRC's custom solution to the

5 See the Utah data transfer process here: https://jobs.utah.gov/edo/udrcadvisory/udrcgovernance.pdf

6 Learn more about the UDRC here: https://udrc.utah.gov/about.html

challenge of data ingestion is the result of a legislatively supported collaboration with the Utah Department of Technology Services.

Connecticut has a SLDS called P20·WIN. However, unlike Utah's centralized data system, P20·WIN is federated. In practical terms, this means that P20·WIN does not warehouse data. Rather, partnering entities oversee control and security of their own administrative data. According to Jan Kiehne, Senior Associate for Decision Support Resources and Connecticut P20·WIN Program Manager, data linkages across entities are only made when an audit or evaluation study request is made and subsequently approved. When a request for data is approved, each entity involved in the request generates a data table that is split into two sections: one section with PII, and one section with analysis data. The respective entity uses a process to generate a random identifier for each individual observation in the data file. The entity then transfers this PII section via an SFTP server to the Connecticut Department of Labor (DOL) while retaining the analysis data file. The DOL receives the PII sections from each entity involved in the data request and completes a probabilistic data match based on the PII. The DOL then destroys all PII, leaving just the randomly generated identifiers. This file is then transferred from the DOL via an SFTP server to the requesting entity or individual, who is also given the analysis data files directly from the involved entities.

## Data Integration

Data integration refers to handling and manipulating data after it has been ingested. Some aspects of ingestion and integration overlap; data de-identification and matching are two examples. Other examples of back-end data integration include procedures, protocols, and practices related to data curation, normalization/ harmonization, transformation, and processing for analysis. In the context of an SLDS, data integration is necessary to produce aligned information that can be used for an analysis of education and workforce data, because partner entities often store information in different formats and use divergent technical definitions.

One SLDS data integration challenge, for example, is how to standardize the data elements from different partnering agencies that are used for matching records. Agencies may capture data on the same underlying variable (e.g., gender), but they may code it differently in their respective repositories. According to Jeremias Solari, Utah currently attempts to ingest data from partners in its native format but serious inconsistencies in reporting have led them to consider adding a data manipulation layer into their ingest process that would ascertain specific variables and manipulate

them to conform to a preassigned data structure. This manipulation layer would complement the filtering service that UDRC already has in place to remove invalid rows (such as null matches or individuals who change their surnames). In addition to ensuring appropriate match processes, creating greater consistency within the data set will help to improve the long-term curation of data and improve analysis.[7]

The P20·WIN approach to data standardization is much different primarily because Connecticut does not maintain a data repository. P20·WIN has a Data Steward Committee that works closely with partnering data agencies to discuss ways in which data elements can be standardized across entities. However, P20·WIN does not have any formal data integration—including standardization—procedures in place. Rather, the responsibilities of standardizing, cleaning, and manipulating the data for analysis fall on the shoulder of the requesting entity or individual researcher at the time of each request.

## Data Security

Data security refers to a wide array of practices, protocols, and procedures focused on guarding against unauthorized access and abusive use of sensitive data. In the context of an SLDS, data security aims to maintain the privacy of students and individuals, in alignment with federal and state statutes that ensure only authorized, approved, or legitimate individuals have access to the data.

Privacy can be addressed using ingestion and integration practices. As mentioned above, both Utah and Connecticut make concerted efforts to de-identify data early into the data ingest process. In the case of Utah, data are encrypted, stripped of PII, matched, and then placed into a UDRC database with restricted, role-based access. The raw PII data are archived in a physical location and protected by exclusive access rights. In the case of P20·WIN, data with PII move securely between the partnering entities and the DOL; all PII are deleted by the DOL immediately after matching.

However, even when states de-identify data, records may still be identifiable. For example, Jeremias Solari noted that individual students may be identified if cell sizes are small in the data and the students live in small, racially/ethnically homogenous areas. Consequently, both Utah and Connecticut set additional security protocols in place. Both states, for example, suppress data values that may lead to unintended secondary

7 J. Solari, personal communication, March 4, 2020.

disclosers. Rather than setting a threshold number as a trigger for suppression, Utah has constructed an algorithm as part of its custom ingest technology solution that detects when reporting values may have a high risk for secondary disclosure. Utah also adheres closely to the special publication of the National Institute of Standards and Technology (NIST 800-53) which catalogues best practices related to security and privacy protections against cyber-attacks, natural disasters, and human error.[8] Connecticut and P20·WIN staff are moving to adopt these standards at the time of this writing.

In addition to protecting student privacy and guarding against any secondary disclosures of PII that may violate state or federal protections, data security practices must also monitor and control data access. UDRC manages data access through a tiered system that stores data as (1) restricted access data, (2) UDRC research-use data, (3) external research-use data, and (4) public-use data (see Figure 3). Restricted access data contain PII and are only accessible to a select number of UDRC personnel. Authorization for access is controlled through the IT Secure Access Management Service (SAMS) system operated by the Utah Department of Workforce Services. UDRC research-use data contain unit-level data records that are de-identified. Access to these data are restricted to UDRC staff and approved external researchers from the partner agencies. External research-use data also contain unit-level records that have been de-identified. These data have also been reviewed and, if necessary, values have been suppressed to protect against disclosures. Public-use data does not contain unit-level data and have been aggregated for public consumption.

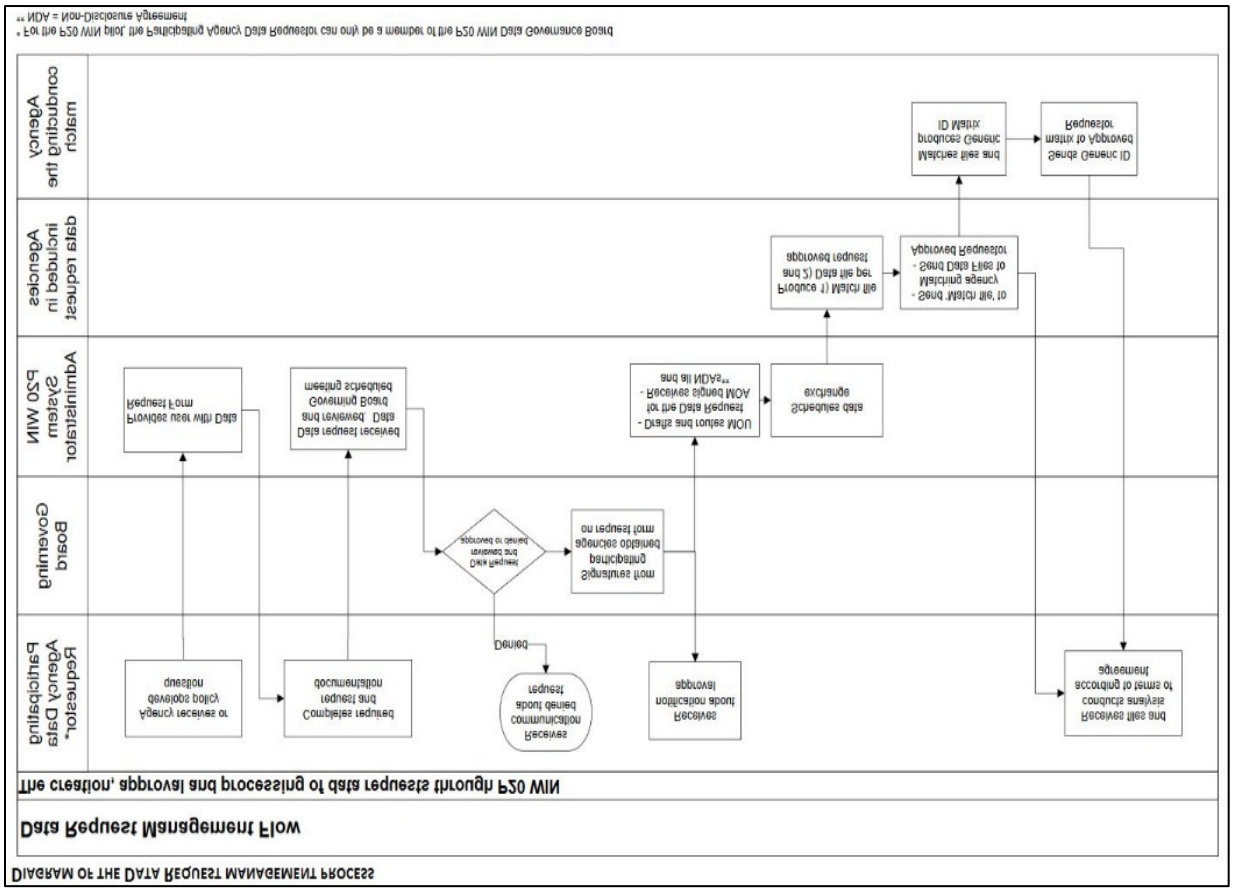
Connecticut does not have an equivalent tiered data access system in place. Again, this is primarily a function of the federated data structure they maintain, in which partner entities are responsible for the security of their own data repositories. Linkages across entities are made only on request for audit or evaluation and the DOL is the only entity (other than the sharing entity) that is ever in possession of PII.

Both Utah and Connecticut make de-identified, unit-level data available to approved external researchers. Each has established processes and business rules to review data request applications, approve requests, make data available, review reports or deliverables, ensure student privacy is protected, and data are destroyed. These

8 Read more about NIST guidelines here: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf

external data request procedures for both states are available online.[9] Figure 4 below illustrates the P20·WIN data request management flow.

Figure 4: Connecticut's P20·WIN Data Request Management Flow



Notable similarities between Utah and Connecticut are that both states involve partner entities in the review and approval of research requests and both states require external researchers to submit any reports they plan to disseminate to the partner entities in addition to the SLDS staff. Notable differences between the two states are that Utah requires an institutional review board (IRB) review. Also, Utah UDRC staff review reports by external researchers (which often includes graduate students) to determine if they are methodologically sound. Finally, Connecticut P20·WIN includes more restrictive language as to who can request data.

9 Read the P20·WIN data request management process here: https://www.ct.edu/files/pdfs/P20WIN-DataRequestProcedure-Final_012020015.pdf; Read the UDRC external data request process here: https://udrc.utah.gov/data-request.html

## Key Questions for the Subcommittee Meeting

At the subcommittee meeting, participants will have the opportunity to consider appropriate ingestion, integration, and security protocols for the California data system. Participants will work in small groups to address these concepts in the context of two of the prioritized use cases: a P20W data set and a data request process.

Those addressing the P20W data set use case will answer the following questions:

1. Based on a proposed data set, which elements are problematic related to quality, frequency, and data security issues?
2. What technical mechanisms should be used to provide the data to a centralized repository? Should there be batch/file-based uploads or real-time uploads? What protocols should be used? If not live, how frequently should data be uploaded?
3. Where should the data be stored? What type of cloud environment could be deployed in phase one?
4. What processes should be used to identify and tag sensitive data? How would the system ensure that such data are secured?
5. What guidelines should attend provisioning access to sensitive data?
6. How would role-based access support security and how could those policies be developed?

Those focused on the data request process will answer these questions:

1. What limits should be put on the data that are available? Which elements are problematic related to quality, frequency, and data sensitivity issues?
2. What technical mechanisms should be used to provide the data when requested? What protocols should be used? What approval process should be used?
3. Where should the data be stored? What type of cloud environment could be deployed in phase one?
4. What processes should be used to identify and tag sensitive data? How would the system ensure that such data are secured?
5. What guidelines should attend provisioning access to sensitive data?
6. How would role-based access support security and how could those policies be developed?

By considering how other states have chosen to structure and assign responsibilities for data transfer processes, California can determine the pros and cons of various approaches.