SLDS Issue Brief

Building a Centralized P-20W Data Warehouse

For over a decade, many states have tried to link data across the P-20W (early learning, K12, postsecondary, and workforce) spectrum, but these efforts were often dominated by a single agency and thus inhibited by political barriers. More recently, some states have succeeded in linking these data by cooperating across multiple sectors to create a new independent agency (or designate an existing one) to be responsible for building and maintaining a centralized P-20W data warehouse. Two such examples are Kentucky and Washington State, whose lessons in building a centralized P-20W data system are shared in this document.

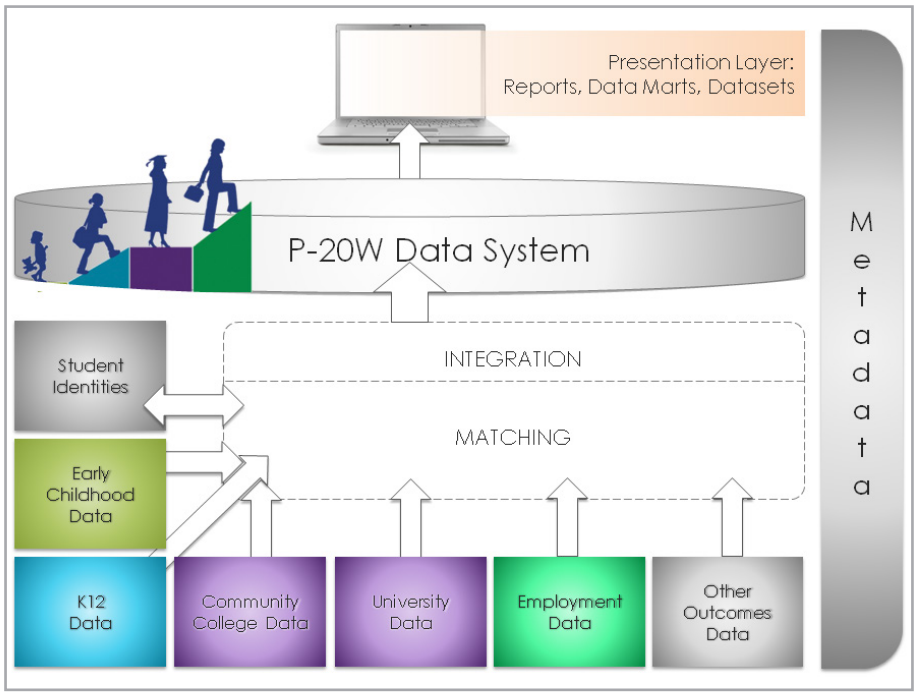

Figure 1. Basic structure of centralized data system

Governance and Administration

In a centralized data system, all participating source systems copy their data to a single, centrally located repository where they are organized, integrated, and stored using a common data standard. As depicted in Figure 1, data in a P-20W centralized state longitudinal data system (SLDS) are periodically matched, integrated, and loaded into a central repository. Users query the system and can access the data that they have been authorized to view and use.

In order to govern a centralized P-20W data warehouse most effectively, a board should be formed consisting of leadership from each agency that provides data. Each source agency may be considered a member of the P-20W agency and still maintain ownership, access, and control of its own data (i.e., no individual agency's data may be accessed without written approval), and all project and fiscal decisions can be agreed upon unanimously. Best practices suggest (and, in some cases, the Family Educational Rights and Privacy Act (FERPA) may *require*) the signing of a memorandum of agreement (MOA) allowing the P-20W agency to access, exchange, and link data, and defining the

This product of the Institute of Education Sciences (IES) was developed with the help of knowledgeable staff from state education agencies and partner organizations. The content of this brief was derived from a Statewide Longitudinal Data Systems (SLDS) Grant Program monthly topical webinar that took place on November 29, 2012. The views expressed do not necessarily represent those of the IES SLDS Grant Program. We thank the following people for their valuable contributions:

Charles McGrew
Kentucky Center for Education & Workforce Statistics

Carol Jenner
State of Washington Education Research & Data Center

Jeff Sellers
SLDS Grant Program, State Support Team

For more information on the IES SLDS Grant Program or for support with system development, please visit http://nces.ed.gov/programs/SLDS.

responsibility of the governance team as data stewards.[1] Additional care may be given to define how the agency works within FERPA. In Kentucky, for example, this was defined in an executive order created by the Governor.

The structure of each state's board varies based on the way the state organizes its sectors. For example, in Washington, the Education Resource Data Center (ERDC), home to the state's unit-record public baccalaureate data system, coordinates with the Department of Labor & Industries, the Office of Superintendent of Public Instruction, the State Board for Community and Technical Colleges, the Employment Security Department, and other agencies to build and maintain the data warehouse. Meanwhile, Kentucky's board includes representation from the Kentucky Education Workforce and Development Cabinet, the Kentucky Department of Education, the Kentucky Council on Postsecondary Education, and the Kentucky Educational Professional Standards Board.

> One of the many benefits of a centralized P-20W agency and its independence from any individual organization is its ability to think about better ways to match, de-identify, and link data, and to identify state-level issues of concern to stakeholders—without being limited by a singular perspective.

Additionally, the P-20W agency should look to include more sectors in the warehouse, and thus on the governance team. Cross-sector data, such as Kentucky's data on educational licensure, allows researchers to ask previously unanswered questions about education and its impact. The inclusion of additional data sources and the building of relationships with these agencies can be overwhelming at first, but are likely to produce information that can be used to improve these individual agencies and education in general.

Under the guidance of the board, the centralized data warehouse is best administered by a full-time staff. The primary role of this staff is to provide information that individual agencies are unable to create—in effect, providing better, timely information for decisionmaking. The staff may also be assigned responsibilities by the state, such as evaluation and assessment of education programs.

Construction and Extraction

Before the centralized data warehouse is built, stakeholders—including researchers, policymakers, legislature, citizens, and other agencies—may be surveyed about their information needs, formatting needs, and any other concerns. Stakeholders may also be asked to identify the data elements that are necessary for research and policy analysis.

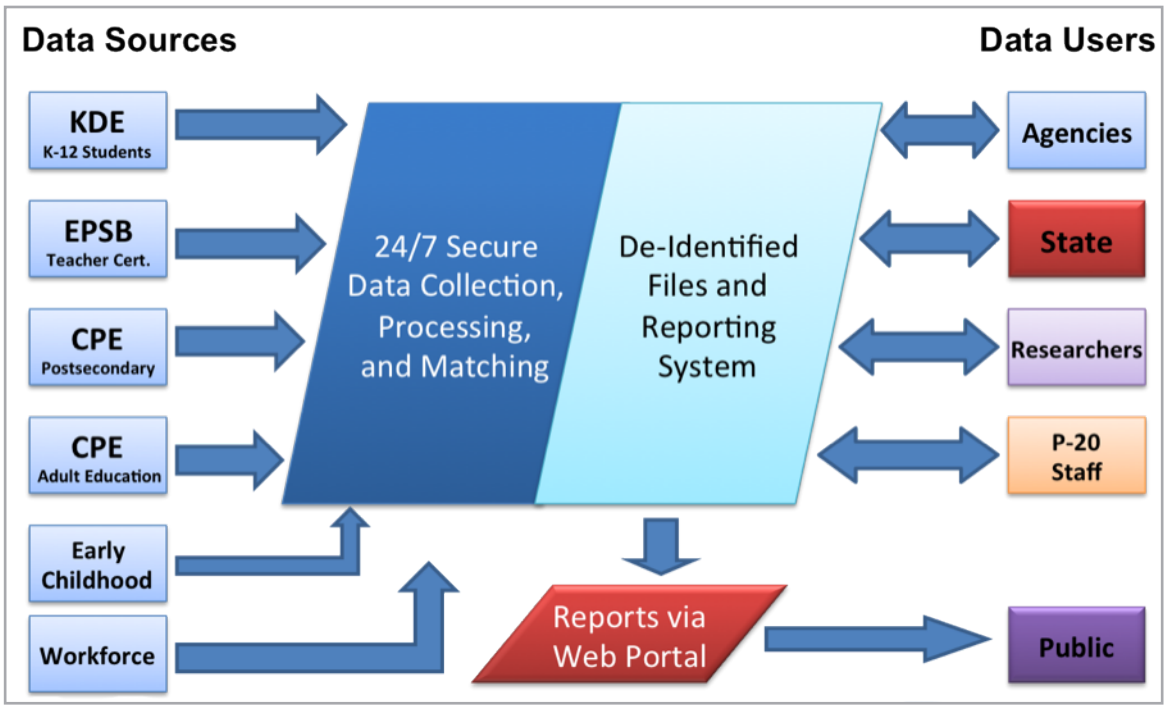

Figure 2. The architecture of Kentucky's longitudinal data system warehouse

[1] As noted, FERPA may require the signing of an MOA, especially in the case of a multi-agency P-20 data warehouse. Visit *http://ptac.ed.gov/sites/default/files/data-sharing-agreement-checklist.pdf* for guidance on this process.

Once the warehouse is built to meet these specifications, the data can be extracted from the systems of individual agencies. One challenge with extracting data is that some source agencies may not have their own warehouse (e.g., early childhood education). Creating the P-20W warehouse architecture from a platform perspective will allow additional data warehouses to be built within the structure for agencies that do not have their own. The platform structure also allows for reporting tools to be leveraged for agencies that have limited reporting capabilities: instead of buying individual enterprise licenses, each agency can access the P-20W tools and use them against their own data or repository. In either scenario, the P-20W agency is not replicating individual agency systems, but facilitating cross-sector work.

Most data warehouses are able to store an unlimited amount of data; however, in order to create efficient systems, only the elements required for P-20W research and policy analysis (as identified by stakeholders) should be collected. Despite unlimited technical capacity, the number of years of data in a P-20W warehouse is limited by the records available in the source systems. For example, Kentucky is able to collect over sixty years of education licensure data, but only four years of K12 data. Similarly, there are rarely technical thresholds to the amount of time that data can be maintained in a centralized warehouse; however, P-20W governance may choose to create a policy around this.

Transformation and Maintenance

Once extracted, data enter the pre-production environment in identifiable formats where they are validated, cleansed, and prepared for matching. Duplicate records are then recognized and individual data are matched across agencies. Identity matching is critical for an effective longitudinal data system. Best practices suggest that in order to improve a system's ability to match, references should be used from outside of the education and workforce sectors (e.g., driver license numbers; last four digits of social security number; and publicly available information such as marriage, divorce, and name change). One of the benefits of a centralized warehouse is that data can be matched and stored, unlike federated warehouses, where data are matched each time a query is made. The storing of matched data allows for the matching process to be ongoing and continually improved, which is an investment in quality data.

Once matched, the data are formatted and may be given P-20W IDs; they are then placed into reporting systems (production environment). If given P-20W IDs, the data exist in de-identified files and personally identifiable information is stored in a separate warehouse with limited access; however, the data are still considered confidential because of their potential for statistical identification in small cell size reports. Therefore, only authorized users and P-20W staff should have access to the de-identified data. Other government and research partners can receive data from the system on an as-needed basis.

ERDC Responsibilities and Principles for Sharing and Using P-20W Data

Principle 1: Education Research and Data Center provides cross-sector, linked data to all data consumers in a consistent, transparent way.

Principle 2: Education Research and Data Center maintains the P-20W data warehouse.

Principle 3: Protecting the privacy of individuals is a priority.

Principle 4: Partner agency data contributors (at the state and local levels) are experts at understanding and explaining the data.

Principle 5: Common understanding and use of data increases its value.

Once in the production environment, the data are ready to be linked and used. However, even after data use, the construction of the data warehouse does not end. Unlike transactional data systems (which are used for operational purposes), longitudinal data systems must be continually improved in order to effectively support decisionmaking.

Additional Resources

Centralized vs. Federated: State Approaches to P-20W Data Systems:
http://nces.ed.gov/programs/slds/pdf/federated_centralized_print.pdf

Education Resource Data Center (ERDC): *http://www.erdc.wa.gov*

Kentucky Executive Order Relating to the Creation of a P-20 Data Collaborative Repository:
http://tinyurl.com/c8o8g34

Privacy Technical Assistance Center (for information on FERPA):
http://ptac.ed.gov